Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 46 (2025), No. 6, pp. 2005–2015
DOI : 10.47974/JIOS-2029

Optimized stacked deep ensemble classifier using hyperparameter tuning for facial expression recognition

Venkata Rami Reddy Chirra *
School of Computer Science Engineering
VIT-AP University
Amaravathi 522241
Andhra Pradesh
India

U. Srinivasulu Reddy †
Department of Computer Applications
National Institute of Technology
Tiruchirappalli 620015
Tamil Nadu
India

K. V. Krishna Kishore §
Department of Computer Science Engineering
Vignan's Foundation for Science, Technology & Research (Deemed to be University)
Guntur 522213
Andhra Pradesh
India

Abstract

Despite significant advancements in FER methodologies, real-time implementation remains challenging due to intra-class variations such as pose, illumination, age, and skin color. These variations cause Single Convolutional Neural Network (CNN) architectures to suffer from high variance, negatively impacting system performance. To address these challenges, this study introduces a method to mitigate intra-class variations by integrating multiple CNN architectures and evaluating them on the FER-2013 dataset. The objective is to develop a Stacked Deep Ensemble Classifier (SDEC) that selects the best-performing models for ensemble learning, reducing variance and enhancing accuracy. The approach leverages both stacking and voting ensemble techniques to improve prediction reliability. Four

* *E-mail:* chvrr58@gmail.com (Corresponding Author)
† *E-mail:* usreddy@nitt.edu
§ *E-mail:* kishorekvk1@yahoo.com

high-performing base models were designed and developed for this work. In the stacking ensemble, the outputs of these base models were combined and processed through a meta-classifier to generate the final prediction. In the voting ensemble, each model independently predicted class probabilities, which were then aggregated, with the final class determined using the argmax function. The proposed models were evaluated on the FER-2013, achieving a test accuracy of 70.5% for the stacking ensemble and 70.4% for the voting ensemble.

Subject Classification: *68T07, 68T42.*

Keywords: *Computer vision, CNN, Voting ensemble, Stacking ensemble, Argmax.*

1. Introduction

Emotions are most commonly communicated through speech, body language, hands, eyes, and facial expressions. According to [1] 55% of emotions are conveyed solely through facial expressions. In a comprehensive study of facial expression analysis [2-3] six fundamental universal emotional expressions have been proposed: anger, joy, sadness, disgust, surprise, and fear. FER has been utilized in the field of HCI [4]. student awareness estimation [5], driver safety [7], and identifying autism disorder patients [9]. Automatic facial emotion recognition systems [10-13] demonstrate strong performance on datasets featuring images captured in controlled environments, but exhibit weaknesses in wild or uncontrolled settings. Conventional techniques, namely PCA (Principal component analysis) [6] LBP [11], and Gabor Filters [7], are not performed well to recognize emotions in an uncontrolled environment due to various challenges in the images. The CNN-based models are reliable and excel at image classification tasks. On large-scale, noisy and wild datasets, the individual CNN model suffers from high variance due to its high variations, which affects the system's performance. Three fully connected (FC) and four convolutional layers were included in the single CNN model [14] for FER and achieved 57.1% accuracy on the FER-2013. Four pretrained

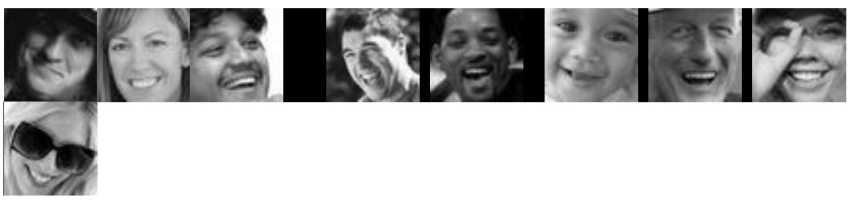

Figure 1

Samples of the FER2013 dataset [21].

models were used in [15] for recognizing emotions from FER-2013. Since 2013, FER using the FER2013 dataset has been a difficult task. It has different variations that include illumination, pose, skin color, aging, and face occlusions. Sample images with different variations of FER-2013 are shown in Figure 1.

In this study, we aimed to reduce high variance by implementing ensemble learning. Unlike single-model approaches, ensemble learning trains multiple models and combines their predictions to achieve superior performance. Initially, various CNN architectures were designed by tuning hyperparameters and thoroughly evaluating the FER-2013 dataset. Multiple model combinations were tested, and four models (Model-1 to Model-4) were selected based on their accuracy. To aggregate predictions, we employed two ensemble techniques: stacking and voting.

This work makes several significant contributions:

- Four CNN-based models have been designed to identify seven distinct facial expressions in real-world environments
- The hyperparameters of the four basic models were fine-tuned to optimize their performance.
- The stacking and voting ensemble techniques were developed to reduce intra-class variations

2. Related Work

In emotional analysis, identifying and categorizing emotions has proven to be a difficult undertaking. Softmax layer with a linear SVM was proposed by [2]. This model demonstrates improved performance across multiple large-scale datasets, such as FER-2013. The model achieved victory in the FER 2013 challenge. It produced 71.2% test accuracy on FER2013. Even though it gives better performance it is limited in terms of execution time. FER using ensemble learning was proposed by [3]. Here, eight individual CNN models were selected based on validation accuracy, and those models are ensembled. This model required more computational resources as eight models were ensembled. A new framework for FER was developed in [4] using the Attentional CNN that used the important regions of the face. The spatial transformer converts important regions of the face into wrapped data. Their proposed model was experimented on CK+, FER2013, FERG, and Jaffe. This model achieved 70.02% accuracy on the FER-2013 dataset. This work failed in addressing how to get the salient

regions when the face has occlusions. Hierarchical committees of deep CNNs were used for facial expression recognition [5]. They have chosen and trained individual CNNs by varying the architectures, normalization, and random weight initialization to form an excellent hierarchical committee of deep CNNs. This work requires more memory due to the hierarchical committee of deep CNNs.

3. Design of various CNN Architectures

During the design of various CNN architectures, we carefully considered multiple hyperparameters, including the number of layers, the count of filters in each convolutional layer, kernel size, pooling mechanism and stride size, the number of FC layers, neurons per FC layer, optimizer, learning rate, activation functions, dropout, batch normalization, and regularization type. Each CNN architecture and its corresponding hyperparameters were fine-tuned based on the accuracy achieved on the FER-2013. We initially experimented with a few convolutional layers and gradually increased their number until we observed an improvement in accuracy, at which point we fixed the number of convolutional and FC layers. Further hyperparameter tuning was performed to optimize the number of filters per layer, pooling method, activation functions, optimizer, FC layer units, and learning rate until performance improvements were observed. After extensive hyperparameter tuning, finalized the design of four optimized CNN architectures. Equation 1 demonstrates the convolution operation between the kernel and the input image in the convolutional layer. As shown in Equation 2, the max-pooling operation selects the highest value within the kernel region, enhancing feature extraction. Equations 3 define the activation function applied after each convolutional layer, which helped accelerate learning and improve generalization. ReLU was specifically chosen to mitigate the vanishing gradient problem by preventing gradient saturation.

$$f_c = \sum_m \sum_n I(m,n)W(i-m, j-n) \tag{1}$$

Here, *f*c represents the feature map, *W* is a convolutional kernel, *I* is input image.

$$f_p = \max i,\ j = 1h, w\ \ x_{i,\ j} \tag{2}$$

The f_p denotes a max-pooling operation feature map.

$$f_e = \{x,\ \text{if}\ x > 0\,\alpha\ (e^x - 1), \quad \text{if}\ x < 0 \tag{3}$$

In this context, f_e denotes feature map of ELU, x signifies an input, and α is a non-linearity parameter.

4. Proposed Ensemble Models

4.1 *Stacking ensemble*

Variations in lighting, pose, age, and skin color in the FER-2013 dataset can lead to suboptimal performance of individual models. To address this issue, the proposed model leverages ensemble learning to reduce variance by integrating multiple models. In this study, a stacking ensemble approach was employed to enhance facial expression recognition using four distinct CNN models. Each model was initially trained on the dataset and generated predictions for the test set. These predictions were then combined and fed into a meta-classifier to produce the final output. The stacking ensemble algorithm is detailed in Table 1, while Figure 2 provides a visual representation of the stacking ensemble framework.

Table 1

Stacking ensemble algorithm

Algorithm Stacking
Input: Train set $T = \{x_i, y_i\}_{i=1}^{n}$
Output: meta- classifier C
Step-1: Train individual models
for k=1 to 4 do
learn $h_{k\,\text{based}}$ on T
end for
Step-2: Generate new data from predictions of individual models
for i=1 to n do
$T_h = \{p_i, y_i\}$, where $p_i = \{h_1(x_i),..,h_4(x_i)\}$
end for
Step-3: Train a meta-classifier
learn C based on T_h
return C

4.2 *Voting ensemble*

Voting approach integrates the predictions of four base models, where each model generates class probability scores. These probabilities

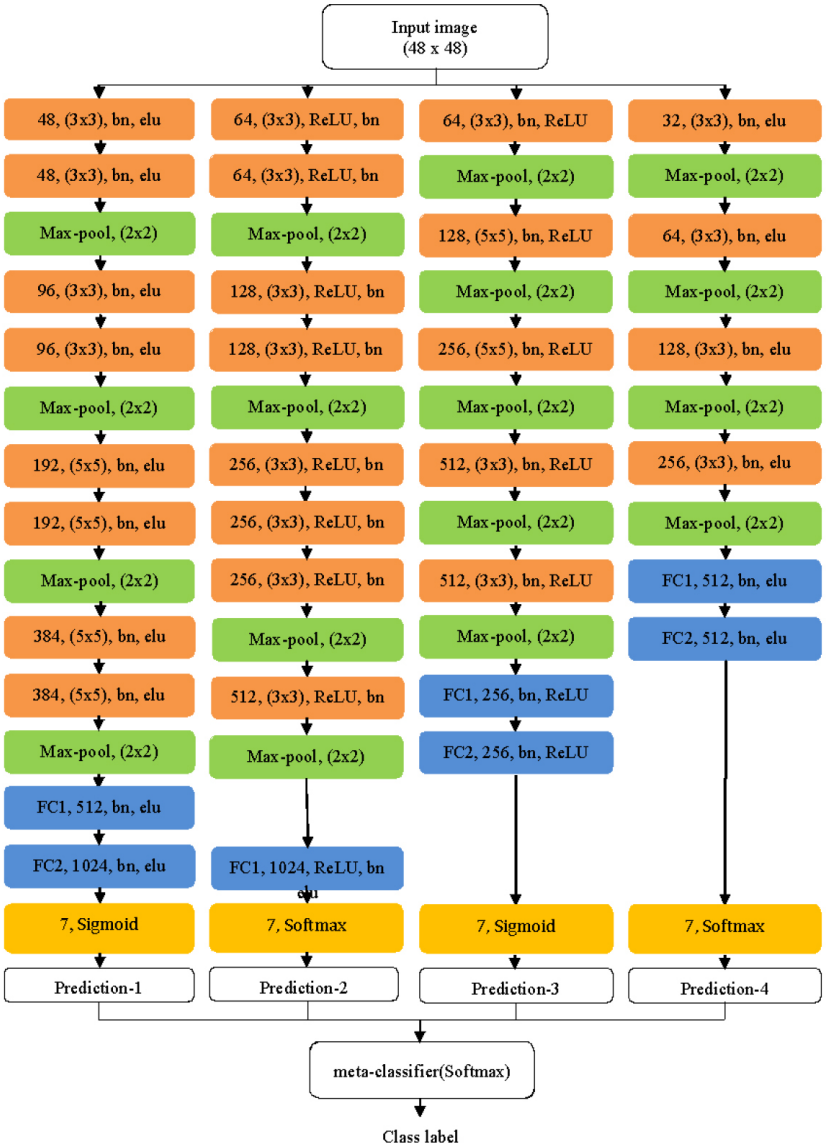

Figure 2

The framework of stacking ensemble

are then aggregated, and the final classification is determined using the argmax function, selecting the class with the highest support. The voting algorithm was provided in Table 2.

Table 2
Algorithm of Voting ensemble

Algorithm Voting
Input: Train set $T_r = \{x_i, y_i\}_{i=1}^{n}$, Test set $T_s = \{X_i, Y_i\}_{i=1}^{m}$
Output: Final class prediction: C_i
Step-1: Train 4 individual models
for k=1 to 4 do
learn h_k based on T_r
end for
Step-2: Generate new data from predictions of individual models
for i=1 to m do
$Y_h = \{P_i\}$, where $Pi = \{h_1(X_i), \ldots, h_4(X_i)\}$
end for
Step-3: Apply Voting technique
$\text{Pred} = \sum_{j=1}^{4} Y_h$
C_i = argmax (Pred)

5. Result And Discussion

The experiments for the proposed study were conducted using the FER-2013. It consists of 35,887 samples with 48×48 dimensions. It has seven labeled expressions. We utilized 80% of samples for training, 10% of samples for validation and testing. The four unique models were trained using the Adam optimizer with a lr of 0.001, a batch size of 128, and a maximum of 150 epochs. Extensive testing was conducted to determine the optimal combination of up to four models based on test accuracy.

Table 3
Individual model accuracies

Model	Accuracy (%)
Model-1	66.59
Model-2	67.21
Model-3	66.20
Model-4	66.47
Stacking ensemble	70.50
Voting ensemble	70.40

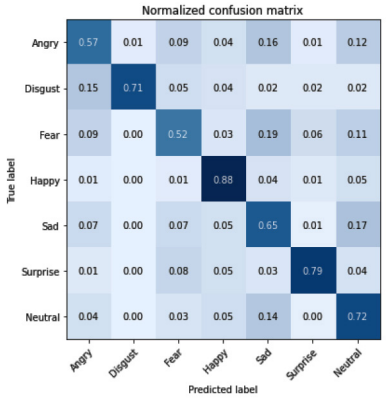

Figure 3

Confusion matrix of Stacking ensemble

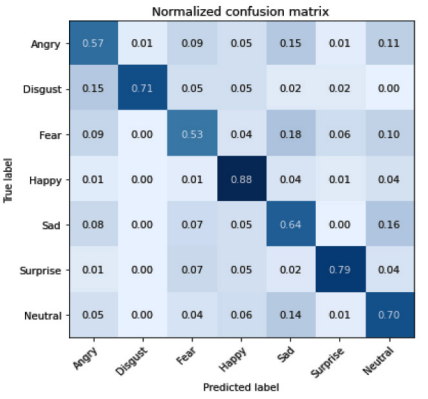

Figure 4

Confusion matrix of Voting ensemble

Individual models demonstrate considerable variability on the FER-2013 dataset, negatively affecting their performance. In the proposed study, multiple sub-models were trained on the same dataset, and their predictions were combined using ensemble techniques. This approach effectively reduced variance and improved accuracy. As a result, both ensemble models achieved accuracies exceeding 70%. The accuracy of the models is presented in Table 3. Additionally, Figure 3 illustrates that the stacking ensemble model successfully recognizes joyful and surprised emotions with 88% and 79% accuracy, respectively. Figure 4 illustrates that the voting ensemble model achieves 88% accuracy in recognizing happy expressions and 79% for surprise expressions, both exceeding the overall model accuracy.

Table 4

Comparison with the state-of-the-art methodologies

Author	Accuracy (%)
Tumen, Soylemez, and Ergen [14]	57.10
Sharma, Patel, and Menaria [15]	64.24
Mohammed et al. [1]	66.40
Chirra, Uyyala, and Kolli [7]	69.10
Shilpa, Kumar, and Jha [8]	68.60
Proposed stacking ensemble	70.50
Proposed voting ensemble	70.40

Table 4 presents the comparison of various existing approaches accuracies on FER-2013. The total time taken by the four models to train the system with the FER-2013 dataset is about 3.7 hours. The Stacking ensemble and voting ensemble methods took 2.85 and 1.53 minutes respectively for testing.

6. Conclusion

The proposed study developed and evaluated an ensemble learning approach for Facial Emotion Recognition on FER-2013. FER-2013 includes various challenges such as variations in lighting, pose, skin tone, age, and facial occlusions. These factors can lead to high variance in individual models, negatively impacting overall system performance. To address this issue and improve accuracy, we implemented ensemble learning by integrating multiple models. Four distinct CNN models were designed, each featuring unique convolutional layer configurations with varying filter sizes and quantities. Model selection was based on the combined testing accuracy of each individual model. Two ensemble strategies stacking and voting were employed to merge these models. In the stacking ensemble, the outputs of the base models were combined and fed into a meta-classifier to generate the final prediction. In the voting ensemble, each model predicted class probabilities, which were then aggregated, with the final class determined using the argmax function. The proposed models were evaluated against existing approaches on the FER-2013 dataset to assess their performance accuracy.

References

[1] T. K. Mohammed, D. N. Vasundhara, S. H. Mehanoor, E. Sreedevi, P. R. Kumar, C. Manihass, and S. F. Baba, "A novel fusion approach for advancement in crime prediction and forecasting using hybridization of ARIMA and recurrent neural networks," *Journal of Information Systems Engineering and Management*, vol. 10, no. 38, pp. 404–420 (2025).

[2] R. Q. Abdulkadhim, H. S. Abdullah, and M. J. Hadi, "Improving cryptocurrency price prediction through advanced LSTM-based deep learning techniques," *Journal of Statistics and Management Systems*, vol. 27, no. 8, pp. 1701–1712 (2024).

[3] B. Shilpa, P. R. Kumar, and R. K. Jha, "LoRa DL: A deep learning model for enhancing the data transmission over LoRa using autoencoder," *J. Supercomput.*, vol. 79, no. 15, pp. 17079–17097 (2023).

[4] N. Arivazhagan, K. Somasundaram, G. B. Mohammad, P. R. Kumar, et al., "Cloud-Internet of Health Things (IoHT) task scheduling using hybrid moth flame optimization with deep neural network algorithm for e-healthcare systems," *Scientific Programming*, vol. 2022, art. 4100352, pp. 1–12 (2022).

[5] P. R. Kumar and B. Shilpa, An IoT-based smart healthcare system with edge intelligence computing, Apple Academic Press, pp. 31–46 (2024).

[6] Sonam and Jyoti, "An Apriori algorithm-based framework for measuring and improving transactional dataset efficiency," *Journal of Statistics and Management Systems*, vol. 28, no. 6, pp. 1179–1189 (2025).

[7] V. R. R. Chirra, S. R. Uyyala, and V. K. K. Kolli, "Deep CNN: A machine learning approach for driver drowsiness detection based on eye state," *Revue d'Intelligence Artificielle*, vol. 33, no. 6, pp. 461–466 (2019).

[8] B. Shilpa, P. R. Kumar, and R. K. Jha, "Spreading factor optimization for interference mitigation in dense indoor LoRa networks," in 2023 IEEE IAS Global Conf. Emerging Technologies (GlobConET), pp. 1–5 (May 2023).

[9] J. Cockburn, M. Bartlett, J. Tanaka, J. Movellan, M. Pierce, and R. Schultz, "SmileMaze: A tutoring system in real-time facial expression perception and production in children with autism spectrum disorder," in Proc. Workshop Facial Bodily Expressions Control Adaptation Games, Art. no. 3 (2008).

[10] C. V. Rami Reddy, K. V. K. Kishore, D. Bhattacharyya, and T. H. Kim, "Multi-feature fusion based facial expression classification using DLBP and DCT," *Int. J. Softw. Eng. Its Appl.*, vol. 8, no. 9, pp. 55–68 (2014).

[11] C. V. Ramireddy and K. V. K. Kishore, "Facial expression classification using Kernel based PCA with fused DCT and GWT features," in 2013 IEEE Int. Conf. Comput. Intell. Comput. Res., Enathi, pp. 1–6 (2013).

[12] P. R. Kumar, "Wireless mobile charger using inductive coupling," *J. Emerg. Technol. Innov. Res.* (JETIR), vol. 5, no. 10, pp. 40–44 (2018).

[13] S. Arvind, B. Unhelkar, S. S. Shankar, P. Chakrabarti, and S. V. Devika, "Advancing cyber threat detection through deep learning in management information systems," *Journal of Discrete Mathematical Sciences and Cryptography,* vol. 27, no. 8, pp. 2409–2417 (2024).

[14] V. Tumen, O. F. Soylemez, and B. Ergen, "Facial emotion recognition on a dataset using convolutional neural network," in 2017 Int. Artif. Intell. Data Process. Symp. (IDAP), Malatya, pp. 1–5 (2017).

[15] S. Sharma, A. K. Patel, and N. Menaria, "Integral transform of product of generalized k-Bessel function and generalized Hurwitz-Lerch Zeta function," *Journal of Interdisciplinary Mathematics,* vol. 27, no. 8, pp. 1757–1764 (2024).

Received December, 2024