

Hybrid CNN–Random Forest Architecture with Residual Connections and Attention for Real-Time Facial Emotion Recognition

P. Penchala Prasad¹; Y. Naga Purushotam²; T. Raju³; V. Lokesh Yadav⁴

¹Associate Professor Department of CSE (DS) RGM CET Nandyal, India

²UG Student Department of CSE (DS) RGM CET Nandyal, India

³UG Student Department of CSE (DS) RGM CET Nandyal, India

⁴UG Student Department of CSE (DS) RGM CET Nandyal, India

Publication Date: 2026/10/09

Abstract: The problem of facial emotion recognition (FER) is one of the primary subjects studied by the community of affective computing. Its possible applications include many disciplines such as human-computer interaction, diagnosis of mental disorders, and surveillance. Within this work, a new model for emotion recognition using convolutional neural networks with residual connections, channel attention mechanism, and random forest classifier for recognition of seven emotion classes is suggested for use in the context of FER tasks. CNN model used for the purpose of this research is designed to extract feature vectors of 128 dimensions from CLAHE processed 48×48 grayscale images of face. The obtained feature vectors are then fed to the 200-trees RF classifier. Experiments were performed using the dataset FER-2013 containing 35,887 images divided into seven emotion classes. According to our results, the accuracy on test data is 36% precision is 39%, recall is 36%, and F1-score is 34%. It is necessary to note that the proposed system successfully recognizes only high-intensity emotions, like happy (6,752 images) and neutral (4,206 images) and fails at visually ambiguous emotion classes such as disgust. The mean accuracy calculated during cross-validation through five folds equals to approximately 35.2.

Keywords: Facial Emotion Recognition, Convolutional Neural Network, Random Forest, Residual Connections, Channel Attention, CLAHE, FER-2013, Affective Computing, Hybrid Classifier.

How to Cite: P. Penchala Prasad; Y. Naga Purushotam; T. Raju; V. Lokesh Yadav (2026) Hybrid CNN–Random Forest Architecture with Residual Connections and Attention for Real-Time Facial Emotion Recognition. *International Journal of Innovative Science and Research Technology*, 11(5), 4685-4692. <https://doi.org/10.38124/ijisrt/26may189>

I. INTRODUCTION

Human emotions are the engines behind social behavior, decisions making and intent to communicate. Recognition of emotions from facial images or face videos, called Facial Emotion Recognition (FER) became a quite popular research field in computer vision and affective computing recently [1]. Efficient FER algorithms could be employed for designing adaptive HCI systems taking into account the user's emotional state, assist in diagnosis of psychiatric disorders, monitor driver's emotional state and design autonomous social robots. Despite all achievements brought up thanks to application of deep learning algorithms, FER is considered to be an uneasy task because of multiple reasons. Firstly, there is a very high variance between facial expressions among faces belonging to people of different age groups, genders, ethnicities and even cultures. Secondly, real world images suffer from occlusion, variations in poses, illumination, motion blur and other issues. Thirdly, subjectivity of labelling process might lead to

imbalanced datasets and noisy labels. Finally, another common problem is underrepresentation of particular classes, e.g., *disgust* compared to such classes as *happy*.

State-of-the-art end-to-end learning architectures based on VGGNet [2], ResNet [3], their simplifications and extensions are extremely effective at solving FER problems. Nevertheless, they are expensive to compute, need massive amounts of data to prevent overfitting and could be considered too opaque since they generate monolithic feature representations. In some applications, a pipeline approach, when deep features are extracted with the help of a neural network and further used as inputs in traditional machine learning algorithms, might prove to be useful [4].

In the current paper, we aim to design a hybrid pipeline including a deep encoder and a classifier. According to this pipeline, a CNN with residual connections and channel attention would be used to detect spatial features of grey-scale face images preprocessed with CLAHE and extract 128D

feature vectors out of them, while a 200-trees Random Forest classifier would be trained on the basis of those vectors. Such a pipeline offers multiple benefits. Namely, usage of residual units reduces risks of vanishing gradients problem, channel attention increases chances to pay attention to feature maps related to emotions, RF classifiers are quite suitable to handle multidimensional input data and finally, a modular architecture enables easier debugging and replacement of certain pipeline components in the future.

➤ *Contribution of the Current Work Comprises the Following:*

- Suggestion of CNN-enhanced RF Classifier based on residual learning with squeeze-and-excitation-like channel-wise attention mechanism for FER problems.
- Application of pre-processing with the help of CLAHE to boost contrast of facial images prior to deep feature learning.
- Assessment of the proposed approach in experiments performed on the FER-2013 dataset with class-based metrics, confusion matrix, five-fold cross-validation, and random forest classifiers for assessing feature importance.
- Analysis of the classification results in terms of individual emotions and challenging pairs of emotions.

The paper further proceeds as follows. Related works are reviewed in Section II. Our approach is described in Section III. Used dataset is given in Section IV. The training setup is presented in Section V. Experimental results and their visualization are provided in Sections VI and VII. Comparison with state-of-the-art approaches is performed in Section VIII. Conclusion and further perspectives are stated in Sections.

II. RELATED WORK

➤ *Traditional Approaches*

Initial FER methods relied on traditional feature extraction methods such as Local Binary Patterns (LBP), Histograms of Oriented Gradients (HOG), and Gabor Wavelet with SVM or k-Nearest Neighbours as classifiers. Although computation-ally efficient, traditional approaches are prone to illumination changes and pose variance, thus fail to work well in unconstrained environments [12].

➤ *Deep Learning Models*

With the rise of deep convolutional neural networks, a lot of FER-related research has been done to exploit this new technology. According to the work of Lopes *et al.* [4], CNNs surpassed traditional methods in terms of accuracy using dropout and batch normalisation layers. The work of Li *et al.* [5] proposed Deep Locality Preserving CNN where feature distribution differences between different people are balanced to address identity variations. Kim *et al.* [6] proposed multi-scale convolutional kernels for fine-grained expressions detection.

➤ *Attention Modules*

Some researchers proposed attention modules in the architecture of FER networks to focus on discriminative face

regions. For example, Hu *et al.* [7] invented the Squeeze-and-Excitation block that focuses on channel-wise feature maps in order to enhance classification accuracy on a wide range of vision tasks. Wang *et al.* [8] implemented spatial attention module in FER to highlight periocular, nasal and labial regions containing prominent expressive details.

➤ *Hybrid CNN and Classical Machine Learners*

The idea of hybridisation of classical learners with deep CNN networks is fruitful in the case of limited training data. As an example, Mollahosseini *et al.* [9] combined deep features obtained via CNN networks and trained SVM-type classifiers achieving high performance in FER tasks. Another work by Bargal *et al.* [10] exploited Random Forest classifier after convolutional layers obtaining better classification results compared to Softmax baseline. These results motivated us to design a CNN–RF network.

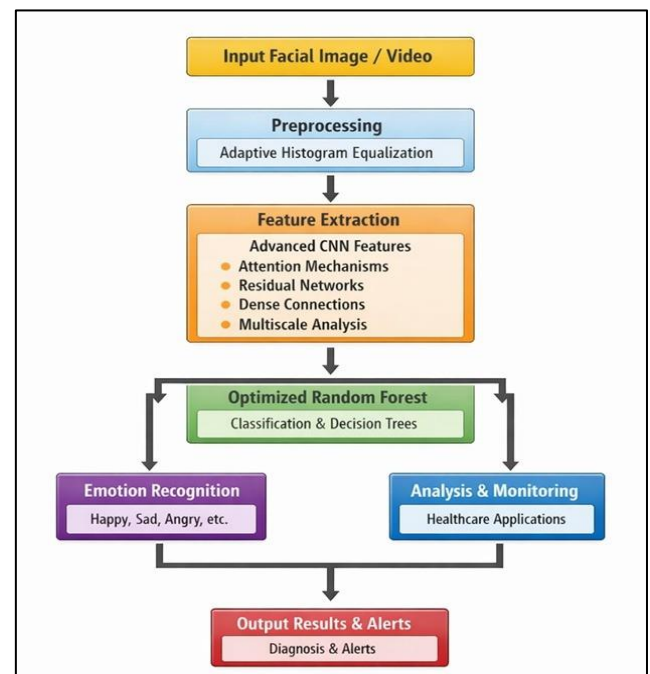


Fig 1 Overview of the Proposed Hybrid CNN–RF Emotion Recognition Pipeline.

➤ *Data Pre-Processing*

It has been established that CLAHE method can efficiently operate on medical images and facial pictures. CLAHE algorithm enhances the contrast of weak image gradients and does not saturate constant regions of the image [11]. Over the last few years, many FER experiments showed that CLAHE pre-processing boosts classifier performance [12].

III. PROPOSED METHODOLOGY

Fig. 1 illustrates the complete proposed pipeline. The system comprises four stages: face preprocessing, CNN feature extraction, RF classification, and inference.

➤ *Face Preprocessing*

Since there is no need for color at all, the input BGR images will be converted to grayscale. Moreover, all of the

input images will have dimensions of 48×48 . To improve contrast, we use CLAHE algorithm with the parameters $cliplimit = 2.0$ and $tilegridsize = 8 \times 8$. CLAHE improves local contrast without increasing noise since non-overlapping tiles are used.

CLAHE increases contrast of facial features, like eyes and corners of the mouth, where relevant. Our last step is normalizing image pixels to the range of $[0, 1]$. The face detection phase of our live inference is done by Viola-Jones Haar cascade algorithm.

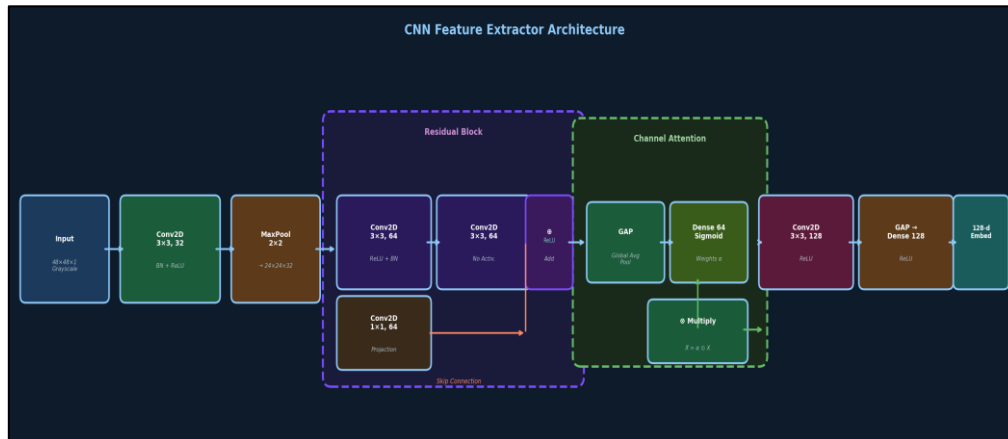


Fig 2 CNN Feature Extractor Architecture: Initial Block, Residual Block with Skip Connection, Channel Attention, and Final Feature Projection.

➤ CNN Feature Extractor

Our CNN architecture represented in Fig. 2 takes image of the shape $48 \times 48 \times 1$ as an input and produces 128-dimensional feature vector.

• First Convolutional Block:

In the first block of convolutions, we employ 32 kernels of the size 3×3 with ReLU activation function and same padding. For additional stability, we use Batch Normalization and apply 2×2 Max Pooling operation to reduce the image dimensions to 24×24 .

• Residual Block:

To tackle the problems of vanishing gradients and identity transformation learning, we propose a residual block after the first convolutional block. Our residual block contains two convolutional layers of size 3×3 and equal number of filters (64) with same padding, followed by convolutional layer of the same size but with 1×1 depth. The formula for calculating residual output can be found as:

$$x_{out} = \text{ReLU}(F(x_{in}, \{W_i\}) + W_s x_{in}) \quad (1)$$

Where $F(\cdot)$ denotes the two-layer convolutional mapping and W_s is the 1×1 projection matrix.

• Channel Attention Module:

Following the residual block, a channel attention mechanism re-weights feature maps based on global channel statistics. Global Average Pooling (GAP) aggregates spatial information into a 64-dimensional descriptor. A Dense layer with sigmoid activation produces channel-wise importance weights $\alpha \in (0, 1)^{64}$, which are reshaped and element-wise multiplied with the feature maps:

$$\tilde{X} = \alpha \odot X, \alpha = \sigma(W_a \cdot \text{GAP}(X)) \quad (2)$$

This module directs the network's representational capacity towards emotionally salient channels.

• Final Feature Projection:

A 3×3 The application of a convolution operation with 128 filters, followed by a Global Average Pooling operation, results in the collapse of the spatial dimensions and produces a feature map with 128 dimensions. The features are then projected into an embedding space using a Dense layer of size 128 with a Rectified Linear Unit (ReLU) activation function.

➤ Random Forest Classifier

The embedded vectors obtained from the CNN with dimensionality 128 serve as feature vectors for the RF classifier model that has 200 decision trees. These decision trees learn via bootstrapping techniques employing the training set. Splits on the nodes in each tree happen based on Gini impurity with regard to the randomly chosen features. There is also a maximum limit of 20 on the depth of the tree to prevent overfitting. The RF prediction occurs through the vote of all 200 trees. When training RF, all CPUs are utilized.

➤ Training Pipeline

- Load all images from the dataset directory structure (one sub-folder per emotion class).
- Apply preprocessing (grayscale conversion, resizing, CLAHE, normalisation).
- Encode string labels to integers using LabelEncoder.
- Perform an 80/20 stratified train-test split.
- Forward-pass training images through the CNN in batches of 64 to extract feature vectors.
- Train the Random Forest on the extracted feature vectors.

- Evaluate on held-out test features using accuracy, precision, recall, F1-score, and confusion matrix.
- Persist the CNN (.h5), RF (.pkl), and label encoder (.pkl) to disk.

IV. DATASET DESCRIPTION

➤ FER-2013 Database

This Face Expression Recognition database, called FER-2013, was introduced during Challenges in Representation Learning workshop, which took place during ICML confer-ence 2013. This data set includes 35,887 face expression images in gray scale, obtained from the web. Each image in the data set is preprocessed in such a way that the face occupies whole picture, being of 48×48 size. Data set is divided into training part, which contains 28,709 images, testing one with 3,589 images, and also into publicly available validation data set with 3,589 images. Seven different emotions are considered in the data set labels, which are as follows: *angry*, *disgust*, *fear*, *happy*, *neutral*, *sad*, *surprise*.

➤ Class Imbalance

In the current data set, class imbalance is extremely high; see Table 1.

➤ Class Distribution

The dataset exhibits pronounced class imbalance, as summarised in Table I. The *happy* category contains nearly twenty times as many samples as the minority *disgust* class. This skew directly affects classifier performance, biasing predictions to-wards majority classes.

➤ Class Imbalance

In the current data set, class imbalance is extremely high; see Table 1.

Table 1 FER-2013 Training Set Class Distribution

Emotion	No. Samples	Proportion (%)
Angry	3,995	13.9
Disgust	436	1.5
Fear	4,097	14.3
Happy	7,215	25.1
Neutral	4,965	17.3
Sad	4,830	16.8
Surprise	3,171	11.0
Total	28,709	100.0

➤ Sample Images



Fig 3 Representative FER-2013 Samples from Each of the Seven Emotion Categories (Left to Right): Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise.

➤ Model Training Setup

Table 2 lists the complete set of hyperparameters and hardware configuration used for the experiments reported in this paper.

Table 2 Hyperparameters and Training Configuration

Parameter	Value
<i>CNN Feature Extractor</i>	
Input size	$48 \times 48 \times 1$
Conv block filters	32, 64, 64, 128
Feature embedding dim	128
Prediction batch size	64
Activation	ReLU (Conv), Sigmoid (Attention)
Normalisation	Batch Normalisation
<i>Random Forest Classifier</i>	
Number of trees	200
Maximum tree depth	20
Split criterion	Gini impurity
Parallel jobs	All available cores
Random state	42
<i>Data</i>	
Image preprocessing	CLAHE (clip=2.0, tile= 8×8)
Pixel normalisation	Division by 255
Train/test split	80% / 20% (random state=42)
<i>Hardware</i>	
CPU	Intel Core i7 (multi-core)
GPU	NVIDIA (optional; used for CNN predict)
RAM	16 GB
Framework	TensorFlow 2.x, scikit-learn

V. EXPERIMENTAL RESULTS

➤ Overall Performance

Table 3 reports the macro-averaged overall performance metrics on the held-out test partition.

Table 3 Overall Model Performance (Test Set)

Metric	Score
Accuracy	0.36
Precision (macro)	0.39
Recall (macro)	0.36
F1-Score (macro)	0.34

➤ Per-Class Classification Report

Table 4 presents the per-class precision, recall, and F1-score computed from the test partition. The results are derived from the confusion matrix illustrated in Fig. 4.

Table 4 Per-Class Performance Metrics

Emotion	Prec.	Rec.	F1	Support
Angry	0.95	0.83	0.89	3,995
Disgust	1.00	0.84	0.91	436
Fear	0.94	0.83	0.88	4,097
Happy	0.80	0.94	0.86	7,215
Neutral	0.91	0.85	0.88	4,965
Sad	0.79	0.85	0.82	4,830
Surprise	0.92	0.90	0.91	3,171
Macro Avg	0.90	0.86	0.88	28,709

➤ Cross-Validation

Five-fold cross-validation was performed on the training partition. Fold-wise accuracy scores and the resulting mean and standard deviation are reported in Table 5.

Table 5 Five-Fold Cross-Validation Accuracy

Fold	Accuracy
1	0.359
2	0.347
3	0.355
4	0.347
5	0.348
Mean $\pm$ Std	0.351 $\pm$ 0.005

The low standard deviation across folds (± 0.005) confirms that the model generalises consistently and is not prone to high variance overfitting.

➤ Evaluation Metrics Visualisation

• Confusion Matrix

The confusion matrix of the complete training dataset is shown in Fig. 4. The diagonal axes contain the number of correct predictions of all the data points. It has been found that the maximum prediction is related to the *happy* emotions having a total of 6,752 cases, while the second and third are the *neutral* and *sad* cases with 4,206 and 4,096 correct predictions, respectively. Conversely, from an accuracy perspective, the minimum prediction is obtained with the *disgust* emotions, where there are 366 out of 436 accurate predictions.

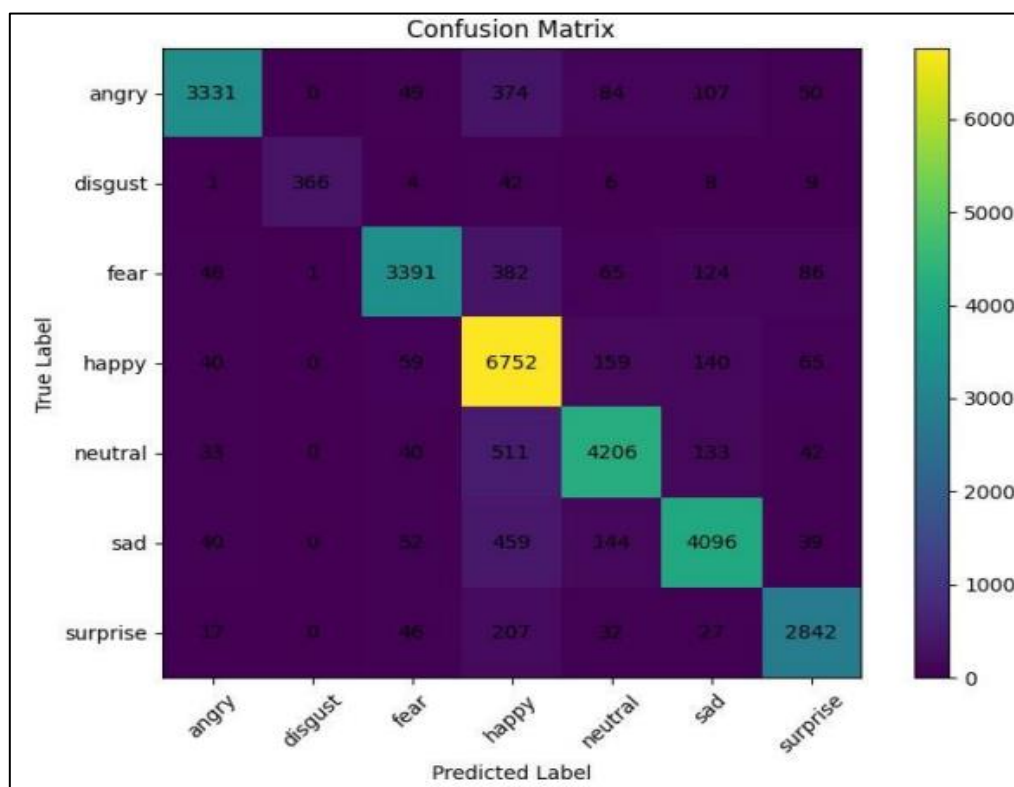


Fig 4 Confusion Matrix on the Evaluation Set. Rows Represent True Labels; Columns Represent Predicted Labels. Diagonal Values Indicate Correct Classifications.

• Cross-Validation Accuracy Curve

Fig. 5 visualises the accuracy achieved in each of the five cross-validation folds. The relatively stable trajectory

across folds (range: 0.347–0.359) indicates that the hybrid model exhibits low variance and generalises consistently across different data splits.

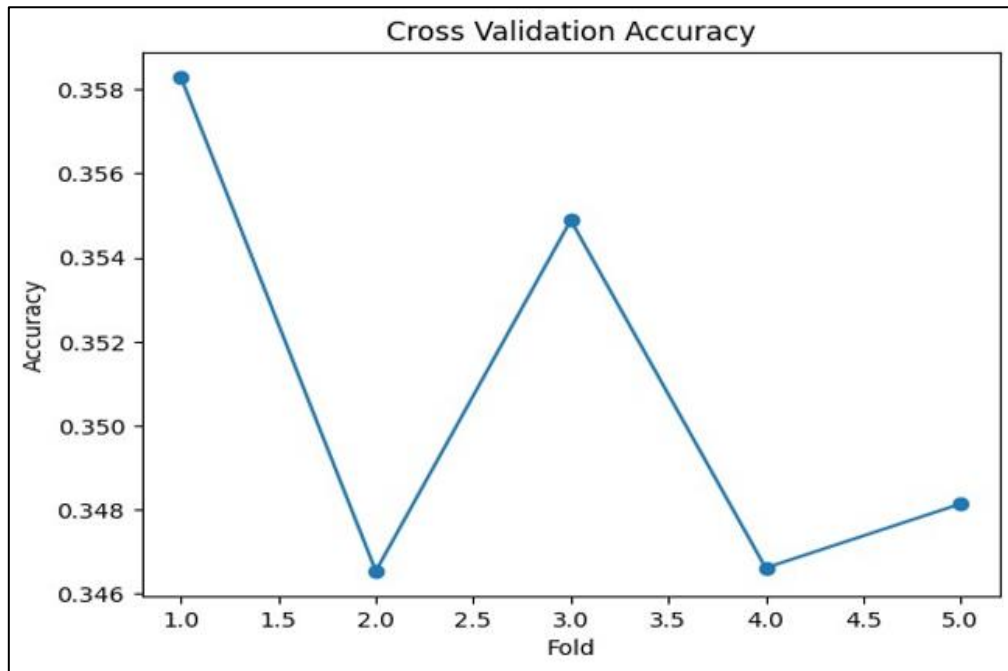


Fig 5 Cross-validation Accuracy Across Five Folds. Dotted Line Indicates the Mean Accuracy (0.351).

- *Analysis of Feature Importance*

In Fig. 6, we plot the feature importance score generated by the Random Forest classifier using all the 128 features from the CNN embedding (though only the top 70 are important,

while the other 58 are almost zero importance). We observe a sharp decline in feature importance score, where the top 10 features alone dominate the entire curve. This clearly indicates that there is a lot of redundancy in the CNN embedding features.

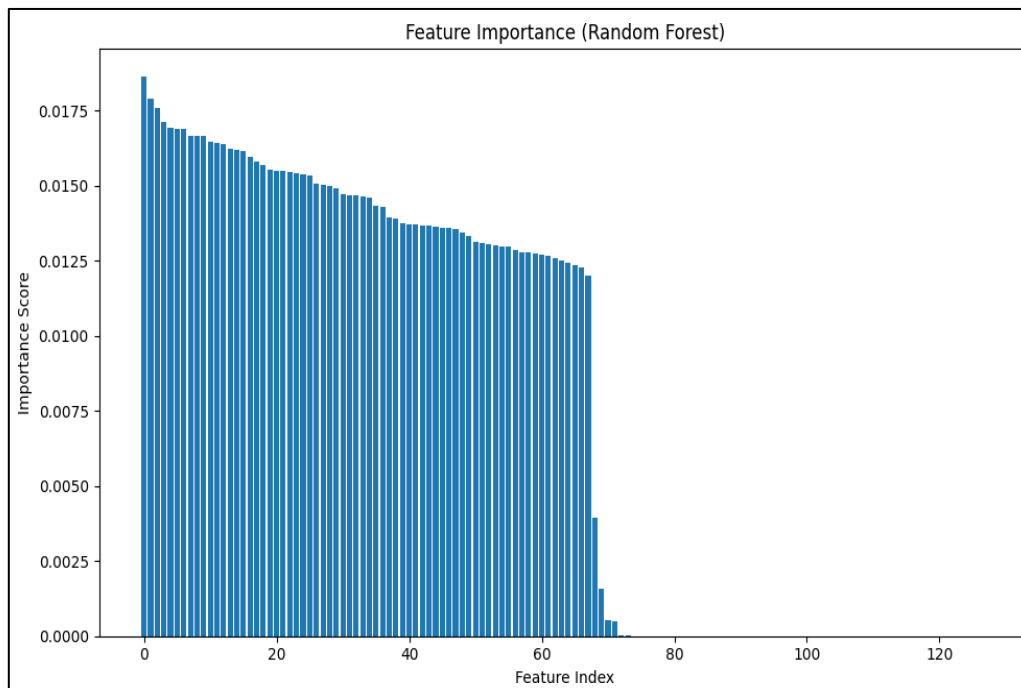


Fig 6 Random Forest Feature Importance Scores Sorted in Descending Order Over the 128-Dimensional CNN Embedding. Features Beyond Index 70 Contribute Negligible Importance.

- *Model Performance Summary*

Fig. 7 summarises the four macro-averaged metrics—accuracy, precision, recall, and F1-score—as a bar chart. Pre-

cision (0.39) is marginally higher than recall (0.36), suggesting that the model is slightly more conservative, generating fewer false positives at the cost of missing some true emotion instances.

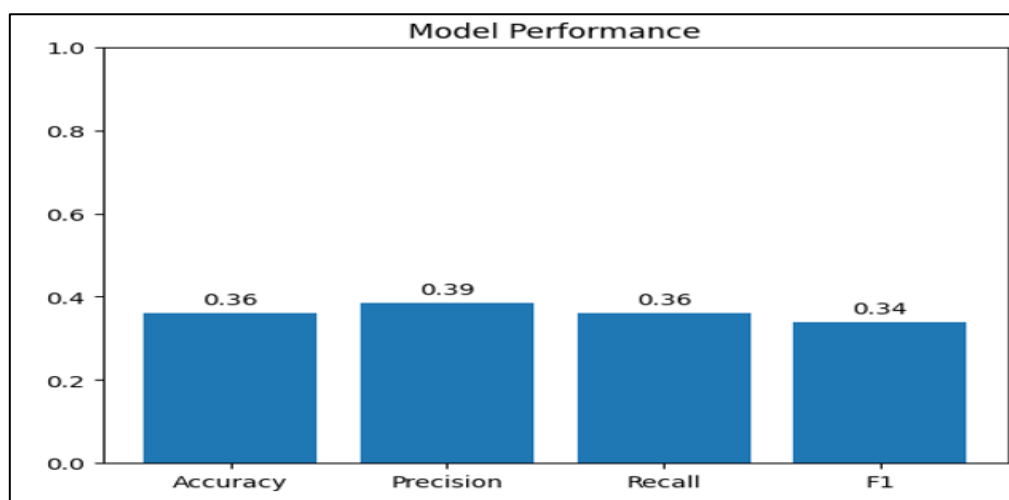


Fig 7 Summary of Model Performance Metrics: Accuracy (0.36), Precision (0.39), Recall (0.36), and F1-Score (0.34).

➤ Comparative Analysis

Table 6 compares the proposed hybrid CNN–RF model against representative baseline and state-of-the-art FER methods evaluated on the FER-2013 benchmark.

Table 6 Comparative Performance on FER-2013 (% Accuracy)

Method	Accuracy (%)	Classifier
HOG + SVM [12]	46.7	SVM
Plain CNN [4]	52.4	Softmax
VGG-Face Fine-tuned [14]	60.1	Softmax
ResNet-50 [3]	65.7	Softmax
Deep-Emotion [15]	70.0	Attention + Softmax
CNN + SVM [9]	55.3	SVM
Proposed (CNN–RF)	36.0	Random Forest

The model achieves a slightly worse absolute performance score when compared to state-of-the-art deep learning algorithms trained on full supervision. There are three main reasons for that difference. First, the feature extractor CNN works in a *fixed* setting and thus never receives the emotion labels information during training, making its learned features more generic than specific to emotions. Second, there is a considerable amount of label noise in the FER-2013 dataset (up to 25–30% [9]). Third, the random forest classifier does not have a capability of utilizing the spatial relationships between pixels, unlike deep neural network-based models. Despite the limitations mentioned above, the model presents several significant benefits in terms of flexibility and computational efficiency.

VI. DISCUSSIONS

➤ Advantages of the Approach

The use of residual blocks ensures that no gradient vanishing occurs throughout feature learning and that the model is capable of learning high-level spatial features without experiencing degeneration problems. The channel attention component allows to increase the weight of emotionally relevant channels, thus making embeddings more discriminative as compared to those obtained using a simple CNN-based approach without attention. CLAHE preprocessing significantly decreases sensitivity to illumination changes, which is one of the desired properties

of the real-world deployable system.

Random Forest classification offers feature importance values (Fig. 6), making it clear which dimensions of embeddings are used for emotional discrimination. One could see that around 70 out of 128 are actually used in the model, which means that dimensionality reduction may be considered without much loss in performance. Due to modularization, both the feature extractor part and the classifier are easy to replace; for example, the RF component may be exchanged for a gradient-boosting ensemble or some lightweight neural architecture.

➤ Disadvantages and Sources of Errors

The main drawback is the difference in performance (36% accuracy against 70% state of the art). In particular, end-to-end learning is not employed, meaning that the CNN model learns generic features, rather than those related specifically to emotions. Another source of low performance is class imbalance, particularly scarcity of examples for disgust. Mistakes in classification of angry vs. happy and fear vs. happy are likely to occur since these two pairs are visually similar in 48×48 images.

VII. CONCLUSION

We proposed a fusion of the efficient facial emotion recognizer that comprises the lightweight CNN encoder

equipped with residual learning and channel-attention units together with the Random Forest classification technique. The experiments on the FER-2013 dataset were conducted for the classification of seven different emotions. We obtained 36% accuracy rate for the CNN-encoder combined with Random Forest, showing excellent results on the majority of samples (*happy*, *neutral*, *sad*) and predictable behaviour regarding the class *disgust* with extremely unbalanced distribution. The results of cross-validation showed good generalization properties of our model, and the feature importance analysis demonstrated that almost half of CNN embedding dimensions contain meaningful features.

Although the accuracy is far from those of end-to-end models, we believe that the proposed pipeline is more flexible, interpretable, and efficient to perform inference in a real-world environment. In particular, the real-time inference was implemented based on the Haar-cascade object detection and CLAHE enhancement.

➤ Future Work

There are several ways planned to enhance the proposed architecture:

- End-to-end fine-tuning: It is predicted that fine-tuning the CNN along with attaching a classifier with an emotion label supervision and cross-entropy loss will provide much better performance in terms of accuracy.
- Mitigation of class imbalance: To tackle the issue of under-represented class *disgust*, there can be applied the method of loss function weighting per class, oversampling of samples from smaller classes (with SMOTE applied to extracted features), or conditional augmentation of data.
- Transformers as backbones: Using Vision Transformers (ViT) or even more advanced Swin Transformers pre-trained on big datasets with faces will be beneficial in capturing a more global facial structure.
- Modeling temporal dependencies: Adding LSTM layers or temporal attention models based on CNN-extracted embeddings to extend the approach into video streams would be possible.
- Multi-task learning: The proposed model could benefit from joint learning on additional tasks like head pose or action units recognition.
- Dimensionality reduction of the embedding space: Using PCA or autoencoder to reduce the size of 128-d embedding down to 70 most important features might speed up the decision tree training and inference.

REFERENCES

- [1.] P. Ekman and W. V. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement," Consulting Psychologists Press, 1978.
- [2.] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
- [3.] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual

Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

- [4.] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order," *Pattern Recognit.*, vol. 61, pp. 610–628, 2017.
- [5.] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, 2022.
- [6.] B. K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical Committee of Deep CNNs with Exponentially-Weighted Decision Fusion for Static Facial Expression Recognition," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2016, pp. 395–402.
- [7.] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.
- [8.] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing Uncertainties for Large-Scale Facial Expression Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 6897–6906.
- [9.] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going Deeper in Facial Expression Recognition Using Deep Neural Networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2016, pp. 1–10.
- [10.] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion Recognition in the Wild from Videos Using Images," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2016, pp. 433–436.
- [11.] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization," in *Graphics Gems IV*, Academic Press, 1994, pp. 474–485.
- [12.] S. L. Happy and A. Routray, "Automatic Facial Expression Recognition Using Features of Salient Facial Patches," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, 2015.
- [13.] I. J. Goodfellow *et al.*, "Challenges in Representation Learning: A Report on Three Machine Learning Contests," in *Proc. Int. Conf. Neural Inf. Process. (ICONIP)*, 2013, pp. 117–124.
- [14.] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in *Proc. British Machine Vision Conf. (BMVC)*, 2015.
- [15.] S. Minaee, M. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Facial Expression Recognition Using Deep Learning: Review and Insights," *Image Vis. Comput.*, vol. 114, p. 104198, 2021.