A Comparative Analysis of CNNs and ResNet50 for Facial Emotion Recognition

Milind Talele
Symbiosis Centre for Research and Innovation, Symbiosis International Deemed University, Pune, Maharashtra, India
milind1248@gmail.com (corresponding author)

Rajashree Jain
Symbiosis Institute of Computer Studies and Research, Symbiosis International Deemed University, Pune, Maharashtra, India
rajashreejain@gmail.com

Received: 6 December 2024 | Revised: 24 December 2024 | Accepted: 4 January 2025

 DOI: https://doi.org/10.48084/etasr.9849

ABSTRACT

Anger, disgust, fear, happiness, sadness, surprise, and neutrality are some of the basic facial emotions that researchers worldwide consider for recognition. Detection of these emotions is important in the present era due to the digital transformation of many processes and human communication. This study analyzes emotion detection methods using the capabilities of deep learning techniques such as a Convolutional Neural Network (CNN) and Residual Network 50 (ResNet-50). The FER2013 benchmark dataset for facial emotion recognition was used for training and testing purposes, along with a few other private images. This study aimed to compare and analyze the performance of the two methods based on several comparative factors, such as architectural differences, feature extraction capability, training dynamics, model performance, computational efficiency, and hardware configuration. The experimental results showed that the ResNet-50 model was significantly more accurate than the CNN, with an accuracy of 85.75% compared to 74%. Although ResNet-50 has higher computational costs, its robustness and accuracy make it the optimal choice for facial emotion recognition tasks. This research provides valuable insight into the capabilities and trade-offs of these models for face emotion recognition techniques.

Keywords-artificial intelligence; basic face emotions; convolution neural network; face emotion detection; deep learning

I. INTRODUCTION

Many studies have focused on human Face Emotion Recognition (FER) using various technologies such as Artificial Intelligence (AI), the Internet of Things (IoT), and electronic sensors. Human body gestures and facial expressions are powerful signals of non-verbal communication. In [1], seven facial expressions were described using the Facial Action Coding System (FACS) which analyzes the movements of the facial muscles, the head, and the eyes. Employing computer-based systems for FER can offer precise and unbiased categorization via trained or pre-trained algorithms. Various fields, such as forensics, medical science, banking, and market research, demand the integration of FER. Customer experience and marketing, human-computer interactions, education, and entertainment can also apply FER systems. In [2], three different directions were proposed for FER: discrete, dimensional, and appraisal-based. Discrete or categorical emotion detection is based on a few basic emotions such as happiness, sadness, anger, etc. Dimension-based recognition includes the Plutchkin emotion wheel [3]. The dimensional approach combines emotion levels at primary, secondary, and other levels, using combinations of basic emotions to provide complex ones. Appraisal-based FER defines emotions as a balanced reaction to events, agents, and objects, and considers balanced reactions to differentiate between emotions and non-emotions.

The FER process involves several steps, such as face detection, feature extraction, and emotion classification. Deep learning has achieved these tasks one step at a time. Deep learning can automate feature extraction without human intervention. Training machine learning models is quicker than Deep Learning (DL), which demands more time, particularly with large datasets, requiring high-level GPU processing. Convolutional Neural Networks (CNNs) and Residual Network 50 (ResNet50) are DL models. Some studies have examined the efficiency of ResNet50 over CNNs [4-5]. This study considered both CNN and ResNet50 for basic FER to compare and analyze their performance. The FER 2013 dataset was used to evaluate the performance of CNN and ResNet-50. Several comparative factors, such as architectural differences, feature extraction capability, training dynamics, model performance, computational efficiency, and hardware configuration, were

considered. The study compared the performance of the models based on accuracy, loss, precision, recall, and F1 scores.

II. PROBLEM STATEMENT AND METHODOLOGY

A critical review of the available literature was carried out based on the evolution of FER, focusing on methods, datasets, and performance. The ImageNet dataset is used to share, search, and interact with images and multimedia content online. [6]. Due to its dimensions, precision, variety, and hierarchical organization, ImageNet served as the foundation for many studies. In [7], the Discrete Fourier Transform (DFT) was used on the most comprehensive video emotive dataset (VideoEmotion8), showing improvements of 51.1-55.60%, performing at a level that was considered state of the art in 2016. In [8], the exceptional image classification efficiency of CNN was demonstrated on the ImageNet dataset. In [9], it was shown that a weakly supervised DL framework can achieve good region-keyword pairs on commonly used norms such as PASCAL VOC and MIT Indoor Scene 67. In [10], a residual learning strategy was presented to facilitate the creation of neural networks that are substantially more extensive than those typically used. An ensemble of these residual networks, with a depth of up to 152 layers, achieved a rate of failure of 3.57% on the ImageNet dataset.

Experiments have been performed on the FER2013 dataset [11] and the private dataset VEMO using the Residual Masking Network (RMN) and the deep Residual Network (ResNet). In [12], four pre-trained models, namely CNN, VGG16, Resnet50, and Senet50, were used with different learning rate values and Adam optimization to recognize basic emotions. VGG16 achieved an accuracy of 97% with a learning rate of 0.001. However, there were some misclassifications, where some emotions of anger and fear were classified as sadness [12]. ResNet50 outperformed the traditional CNN model in terms of accuracy, F1 score, recall, and precision. These results indicate that the added depth and enhanced feature extraction capabilities of the ResNet50 architecture contribute to its superior performance in facial emotion recognition tasks compared to standard CNN models [13-15].

In [16], Interval Type-2 Fuzzy Logic (IT2FL) was used to handle uncertainties in facial analysis. In [17], a neuro-fuzzy inference system, optimized through particle swarm optimization techniques, was used to overcome the shortcomings of traditional methods. In [18], YOLOv8 was used for FER. YOLOv8 was trained and tested on a dataset containing images of faces expressing seven basic emotions, achieving a mean average precision of 0.837 across all emotion classes, demonstrating the model's effectiveness. The results also highlighted potential areas for improvement, such as addressing dataset imbalances and refining the model's ability to detect smaller or off-center facial expressions. Basic facial emotions can form complex emotions [19].

In [20, 21], ResNet-50 outperformed basic CNN architectures in terms of accuracy and model performance. This study investigates this inference by conducting a comprehensive evaluation of ResNet-50 across multiple datasets, including FER2013 and the author's own images. This study systematically compares both models and evaluates them based on computational efficiency. Most previous studies, such as [22, 23], focused on single-model approaches or transfer learning from pre-trained networks. This study performs a systematic comparison of CNN and RestNet-50, focusing more effectively on relevant facial features, performance metrics, and processing time. This study explores various architectural aspects, with CNNs being widely used for image processing and ResNet-50 employing residual learning to address the vanishing gradient problem. The depth, complexity, and feature extraction capabilities of ResNet-50, along with its training dynamics, are analyzed to determine their impact on FER. Furthermore, the performance of these models was evaluated using key metrics, such as F1 score, recall, accuracy, and precision, while also considering computational efficiency in terms of training time and inference speeds. Table I describes the differences between the two models. [24-25].

TABLE I. MODEL ARCHITECTURES

Feature	CNN	ResNet-50
Architecture	A basic stack of convolutional, pooling, and fully connected layers.	Incorporates residual learning with skip connections to prevent vanishing gradients and enable deeper networks.
Depth	Limited depth (shallow networks) due to vanishing gradient issues in deeper models. This CNN had three layers.	Very deep (50 layers).
Building blocks	Convolution → ReLU → Pooling → Fully connected.	Residual blocks with bottleneck layers (1×1, 3×3, and 1×1 convolutions), batch normalization, and skip connections.
Gradient flow	Prone to vanishing gradients in deep networks, limiting depth.	Skip connections enable gradients to flow through the network more effectively, allowing deeper architectures.
Model complexity	Relatively simple and easy to implement.	More complex, with added layers, skip connections, and bottleneck designs requiring careful implementation and tuning.
Accuracy	Moderate performance for tasks such as image classification and emotion recognition.	Higher accuracy due to better feature learning and ability to train deeper models on large datasets.
Training time	Faster due to fewer layers and simpler design.	Longer training time due to increased depth, although training is more stable.
Flexibility	Suitable for smaller datasets or simpler tasks.	More suitable for large-scale datasets and complex tasks due to its robust design.
Overfitting	Higher risk of overfitting in deeper versions without additional mechanisms.	Reduced overfitting due to regularization from residual learning and batch normalization.
Parameter Efficiency	Fewer parameters may lead to lower representational power for complex tasks.	The bottleneck structure reduces the number of parameters while preserving performance, making it efficient for its depth.

A. Convolution Neural Network (CNN)

CNN is an advanced feedforward neural network architecture capable of generating informative feature representations from input images. It leverages non-linear

activation functions to enhance learning and classification capabilities. CNN is one of the deep learning network architectures useful for FER. From the input layer to the output, it has interconnected multiple layers, such as convolution, pooling and dense layers, before it provides a decision. The convolution layer extracts features to produce feature maps. An activation layer between the convolution and pooling layers offers nonlinearity to the network. The activation functions are applied element-wise to the output of the convolution layer. Some functions, such as ReLU and tanh, are useful for this purpose [26-27]. The pooling layer supplies filters that can independently identify essential features and reduce the size of the data. This helps in CPU and memory usage. To handle large image sizes efficiently, CNN must have a sufficient amount of training data and GPU processing.

A hierarchy model builds a network where every neuron is connected. The resulting feature maps are flattened and multi-dimensional image arrays are transformed into a single-sized array for processing by the fully connected layers, such as ANN. Haar cascade was used to detect the face in the Area of Region (AOR) [28-29]. Figure 1 shows a CNN block diagram, Figure 2 shows the procedure employed for FER using CNN, and Table II describes its parameters.

TABLE II. CNN PARAMETERS

Parameters	Formula	Experimental values considered
Input feature map	P	32
Output feature map	Q	64
Input filter	$R \times S$	3×3
First convolution output	$R \times S \times Q \times P$	3×3×64×32
Output with bias across each feature map	$(R \times S \times P+1) \times Q$	(3×3×32 +1)×64

The size of each output CNN layer is calculated by:

$$O_L = I_s - (F_S - 1) \quad (1)$$

where O_L is the output layer, I_s is the input size, and F_S is the filter size. Padding is calculated by:

$$Padding\ (P_S) = I_s + 2 * (F_S - 1) \quad (2)$$

where P_S is the padding size. The CNN block diagram shows the process of FER in stages.

1) Convolution

- Input Layer: The system starts with an input image of a face. Typically, this image is preprocessed to grayscale and resized to ensure uniformity across the dataset.
- Convolutional layers with ReLU activation: In the first convolutional step, the network extracts key facial features, such as the eyes, nose, and mouth, using convolutional filters. The activation function applied here is ReLU (Rectified Linear Unit), which introduces nonlinearity by replacing all negative pixel values with zero.
- Pooling layers: Pooling reduces the spatial dimensions of the feature maps while retaining the most important features. This simplifies computations and minimizes the risk of overfitting. Usually, max pooling is used, which selects the maximum value in each region of the feature map.
- Deeper convolutional layers: Additional convolutional and pooling layers are used to extract more complex and abstract features, such as the overall structure of the face.

2) Classification

- Flattening: The final feature maps are flattened into a one-dimensional vector to prepare the data for input into the fully connected layer, which is input to an ANN.
- Fully connected layer: The flattened vector is passed to a dense layer, where every node is connected to each other node in the subsequent layer. This layer acts as a classifier that maps the learned features to different emotion categories.
- Output layer (emotion categories): The output layer assigns probabilities to predefined emotion categories (Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise). The category with the highest probability is the predicted emotion.

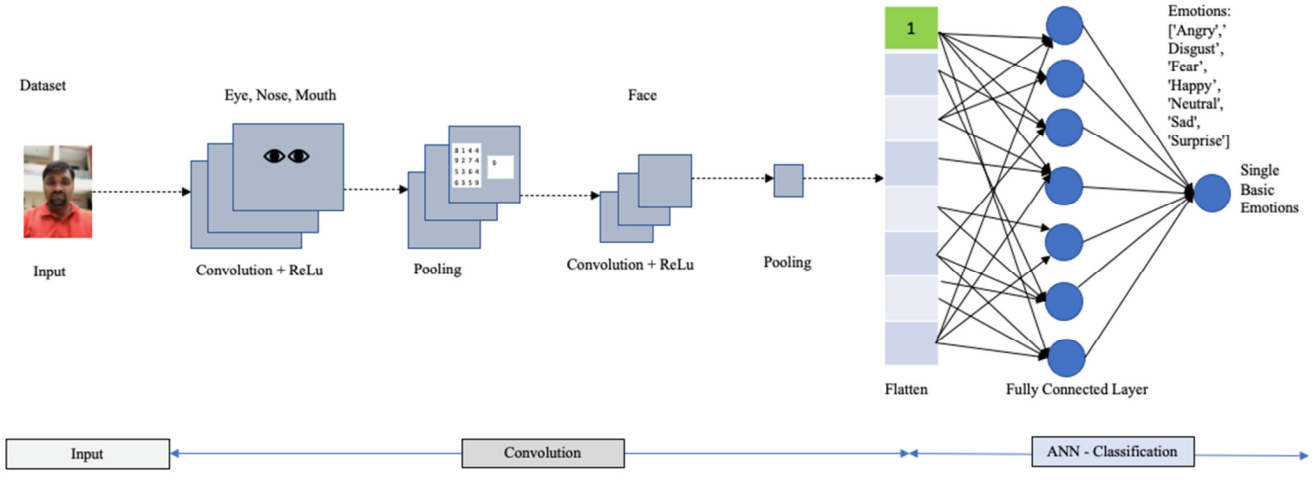

Fig. 1. CNN block diagram.

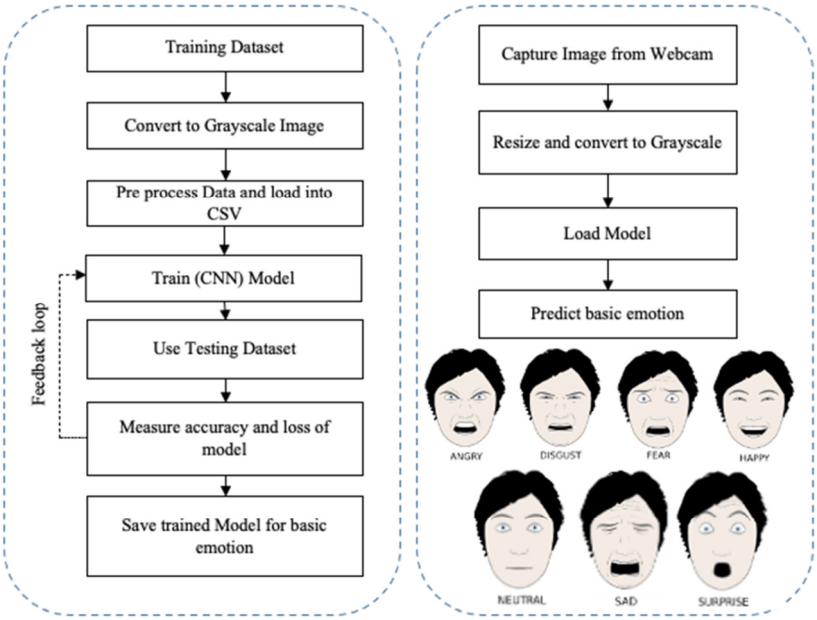

Fig. 2. Flowchart of the FER procedure using CNN.

B. ResNet-50

ResNet-50 is a pre-trained model with 50 layers and a deep neural network architecture. It has been widely adopted in different computer vision tasks due to its ability to capture intricate image features. ResNet-50 addresses the issue of vanishing gradients in complex CNNs. In such networks, as the architecture becomes deeper and more complex, the learning rate decreases leading to a rise in training accuracy but a drop in validation accuracy after a few convolutional blocks. During backpropagation, the gradients diminish, making it harder to adjust the weights effectively. ResNet-50 introduces skip connections and does not change the output shape. These connections ensure better gradient flow, enabling deeper networks to learn efficiently. By using ReLU activation functions, ResNet-50 facilitates learning in highly complex systems. The 1×1 skip connections match the input shape, ensuring smooth integration without any gain or loss in output dimensions. Skip connections improve the validation score as the network continues to learn, making it highly effective for deeper architectures. ResNet-50 has the following layers:

- Stage (1): This stage accepts the input and performs initial convolution and pooling to reduce the size of the input image array. This stage also extracts the basic features.
- Residual stages (2-3): These are important stages for computation where multiple residual blocks consist of two convolutional layers with batch normalization and ReLU activation. These layers connect with the input layer through a shortcut connection to mitigate the vanishing gradient problem by allowing the network to learn residual functions instead of learning the selected mapping.
- Output stage: The last residual stage passes through pooling and fully connected layer for emotion classification.

Figure 2 describes the ResNet-50 process flow. The logic in the residual network is to learn from the output Y, which is a function of the input X:

$$Y = F(X) + X \quad (3)$$

where X is the input identity mapping, Y is the output layer, and $F(X)$ is the learning residual function. However, ResNets shift the focus to learning the function $F(X)$, which represents the change needed to transform X into Y. This is because when $F(X) = 0$, the output Y directly equals the input X.

This approach addresses the problem of vanishing gradients often encountered in deeper networks by using skip connections. Skip connections help by adding the input X directly to the output of a layer block, effectively allowing the network to learn the residual changes $F(X)$ instead of the full output. By targeting $F(X)$ to be zero, ResNets facilitate the learning process, as shown in Figure 3. The parameters used for experimentation for ResNet-50 are listed in Table III.

TABLE III. PARAMETERS OF RESNET-50

Parameter	Value/calculation	Description
Input (X)	48×48×C	Input size where C is the number of channels (e.g., 1 for grayscale images like FER2013).
First weight layer	Kernel size: 3×3, Stride: 1, Padding: 1, Channels: C_{out}	A convolutional layer. C_{out} Is the number of output channels, e.g., 64 in ResNet50
Output of first layer	48×48×C_{out}	Output size is maintained by using padding.
Activation x $F(X)$	ReLU or another non-linear activation is applied to the output of the first weight layer.	Activates the non-linear features.
Second weight layer	Kernel size: 3×3 times 3×3, Stride: 1, Padding: 1, Channels: C_{out}	Another convolutional layer producing an output of the same size.

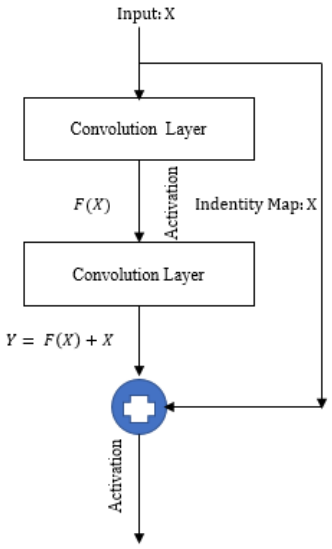

Fig. 3. ResNet architecture.

C. FER2013 Dataset

The FER 2013 dataset was utilized to evaluate the performance of the CNN and ResNet-50 models. It is a large dataset with 36,321 images for robust training and testing in all seven basic emotion categories. All images are grayscale for lower computational complexity. The dataset has been widely used, as it is open-source and freely accessible. Table IV describes the FER 2013 dataset. The dataset was divided into 80% for training and 20% for testing. The preprocessing steps included cleaning the data, identifying the Region Of Interest (ROI), converting images to grayscale, removing blurred or unclear images, and resizing them to 48×48 pixels. This dataset has unbalanced emotion classes and lacks diversity.

TABLE IV. FER2023 DATASET

Emotion categories	Training images	Testing images	Total images
surprise	3236	828	4064
sad	4969	1170	6139
neutral	5013	1247	6260
happy	7195	1856	9051
fear	4134	1049	5183
disgust	467	142	609
angry	4024	991	5015
Total	29038	7283	36321
Split ratio	80%	20%	

III. RESULTS

This study aimed to compare and analyze the performance of CNN and ResNet50 in FER based on architecture, training, testing, hardware utilization, and performance.

A. Model Architecture of CNN and ResNet-50

Although CNNs work feedforward, the Resnet50 feeds the results back into the network. The residual units in ResNet50 enable deep layers to learn from the shallow layers. The total number of parameters is almost 23 million in ResNet50 compared to 1 million in CNN.

B. Training CNN and ResNet50

FER2013 was used for training. Sample results were captured, as shown in Tables V and VI for CNN and ResNet50, respectively. These tables show epoch, time per step, and training and validation accuracy and loss during training and validation.

- Epoch: The epoch number, ranging from 1 to 60 (CNN) and 1 to 48 (Resnet50), indicates the training cycle.
- Step: The number of training steps completed during the epoch.
- Time per step: The average time taken for each training step in seconds (s).
- Loss: The training loss, which measures the degree to which the model's predictions align with the real values.
- Accuracy: The training accuracy indicates the percentage of accurate predictions made by the model. Higher values are better.
- Validation loss: The loss during validation mirrors that of the training loss. It is computed on a distinct validation dataset.
- Validation accuracy: The validation accuracy shows the percentage of accurate predictions on the validation or test dataset.

Accuracy consistently increased in both models for both training and validation. A training accuracy of 74.45% was obtained in CNN against 85.71% in ResNet50. Resnet50 learns from early activation layers to deeper into the network, converging much faster than CNN. Figure 4 indicates the relationship of training accuracy and loss epoch-wise for CNN, whereas Figure 5 indicates epoch-wise accuracy, loss, precision, and F1 scores for ResNet50.

TABLE V. FER TRAINING USING CNN

Epoch	Time per step (s)	Training loss	Training accuracy	Validation loss	Validation accuracy
1/48	780	2.121	0.209	1.693	0.3361
20/48	142	1.1167	0.5766	1.0999	0.5858
48/48	145	0.684	0.74459	1.0321	0.6544

TABLE VI. FER TRAINING USING RESNET50

Epoch	Time per step (s)	Training loss	Training accuracy	Validation loss	Validation accuracy
1/60	541	1.9926	0.8535	1.8343	0.8571
35/60	428	1.5432	0.8571	1.6395	0.8571
57/60	423	1.4024	0.8571	1.567	0.8571
60/60	423	1.386	0.8571	1.5588	0.8571

Precision, recall, and F1 scores were evaluated using:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (4)$$

$$Recall = \frac{True\ Positive}{True\ Positive + Flase\ Negative} \quad (5)$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision \times Recall} \quad (6)$$

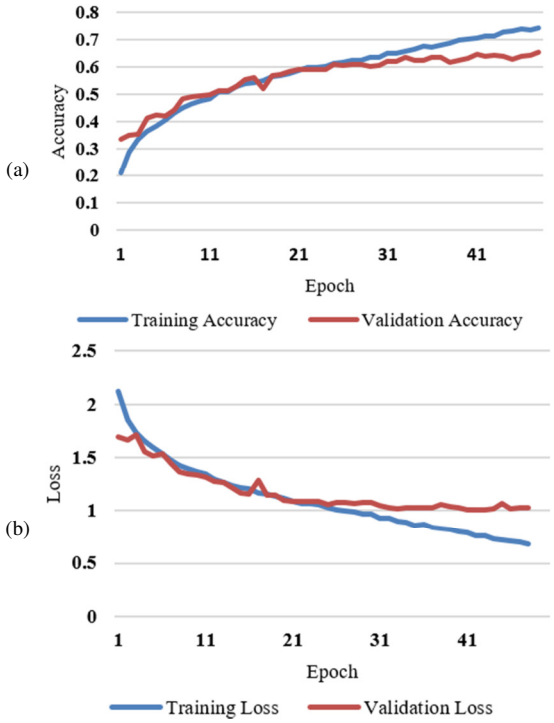

Fig. 4. (a) Accuracy and (b) Loss for CNN.

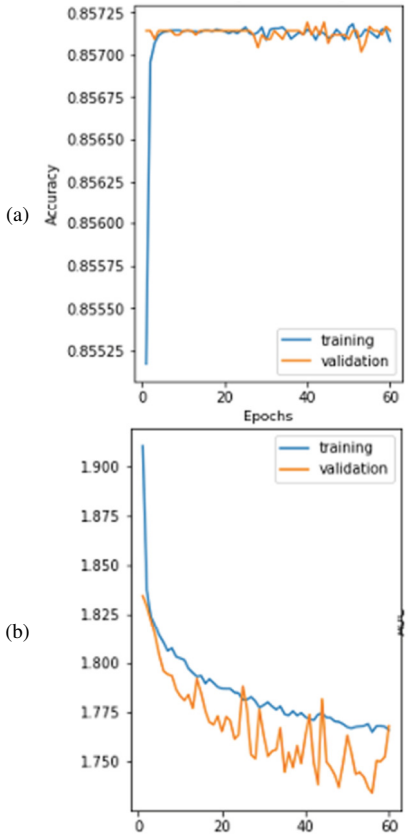

Fig. 5. (a) Accuracy and (b) Loss for ResNet-50.

The accuracy for the training set starts high and remains relatively stable at around 0.85. The training loss decreases steadily, suggesting that the model is learning and improving its performance on the training set less. Table VII presents the hyperparameters chosen. The training and validation losses were monitored to ensure the models were learning correctly, and adjustments were made to obtain the best performance.

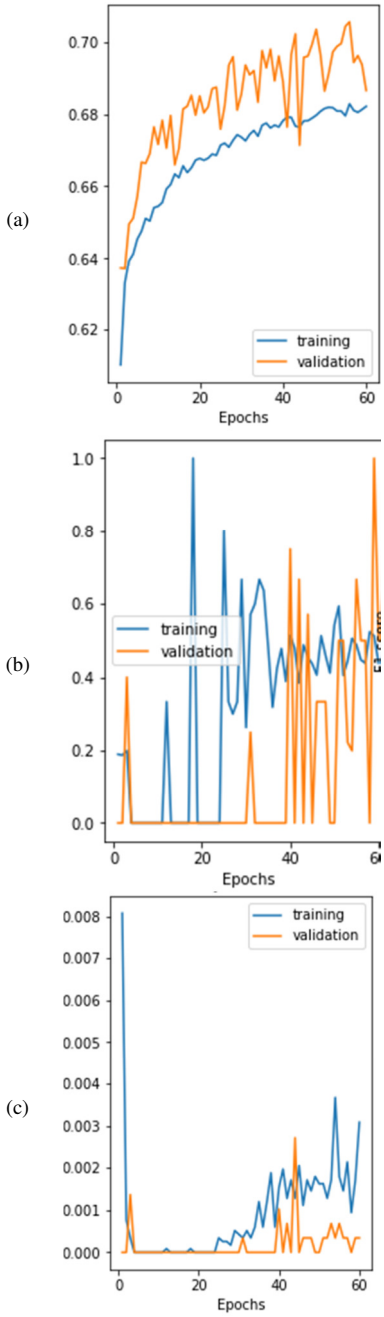

Fig. 6. (a) AUC, (b) Precision, and (c) F1-score for ResNet-50.

TABLE VII. HYPERPARAMETERS USED

Hyperparameter	CNN	ResNet-50
Learning rate	0.001 (Adam)	0.001 (Adam)
Batch size	128	64
Epochs	48	60
Class mode	Categorical	Categorical
Input shape (Grayscale)	(48, 48, 1)	(48, 48, 1)
Optimizer	Adam	Adam

ResNet50 shows consistent and stable accuracy across epochs, as the training and validation accuracies do not fluctuate significantly, indicating better stability and generalization from the beginning. CNN shows a gradual improvement in accuracy, indicating a steady learning process. For ResNet-50, each epoch took approximately 135-150 s per step as a deeper architecture, whereas the CNN epoch takes around 32-33 s per step. This shorter time per epoch is due to the simpler architecture and fewer layers to process.

C. RAM, GPU, and CPU Utilization

Tables VIII and IX show CPU, GPU, memory, and storage usage at different epochs. For CNN, the total memory of the system was 15.1 GB. The percentage of CPU and GPU usage was increasing as the training process was nearing completion. The usage of the CNN model on mobile phones was more difficult, mainly due to limitations on limited processing speed and phone memory, especially when it comes to real-time face identification and recognition using images or videos.

TABLE VIII. CNN: CPU, RAM, GPU, AND DISK UTILIZATION

Epoch	CPU (%) utilization	GPU (%) utilization	RAM (GB)	Disk (GB)
1	11	1	1.8	2.5
4	77	14	1.8	2.5
8	77	14	1.8	2.5
16	76	19	1.8	2.5
20	76	16	1.8	2.5
42	58	19	1.8	2.5
46	76	29	1.8	2.5
48	78	29	1.8	2.5

TABLE IX. RESNET-50: CPU, RAM, GPU, AND DISK UTILIZATION

Epoch	CPU (%) utilization	GPU (%) utilization	RAM (GB)	Disk (GB)
1	68	20	0.4	4.1
4	79	31	1.1	4.5
10	79	31	1.1	4.9
17	80	32	1.1	5.7
19	80	32	1.1	5.9

D. Model Testing

The model was loaded on a Mac for testing with a training Kaggle kernel. Table X shows the machine's specifications. A web camera was used to capture self-images. Figure 7 shows the procedure followed. Images captured through the webcam were converted to grayscale for efficient processing. ROI was converted into an array on which the trained classifiers were adjusted until the required accuracy was achieved and the model was saved.

TABLE X. MACHINE SPECIFICATIONS

Details	Testing-mac machine	Training Kaggle kernel
Disk capacity	120 GB	2.6GB
Memory	4 GB 1600 MHz DDR	15.2GB
CPU	4 GHz Intel Core i5	Xeon
GPU	Intel HD Graphics 5000,1536 MB	NVidia K80

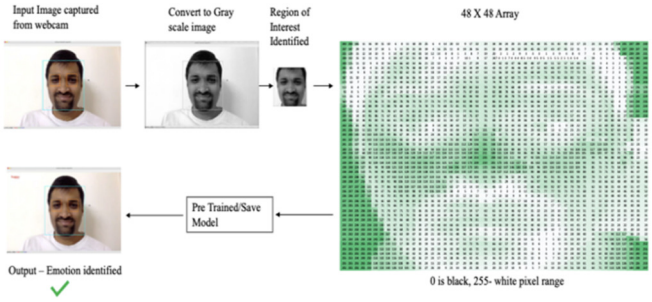

Fig. 7. FER model test with new input image with CNN.

E. Comparison of Model CNN and ResNet50

Table X shows a performance comparison between CNN and ResNet-50.

TABLE XI. PERFORMANCE OF CNN AND RESNET-50

Metric	CNN	ResNet-50
Accuracy	74 %	85.75%
Loss	0.68 %	1.72 %
Time per step range (s)	(142-145)	(278-280)
Val. accuracy	65 %	85.8%
Val. loss	1.03 %	1.69 %
Dataset	FER 2013	FER 2013
Precision	0.89	0.94
Recall	0.86	0.92
F1-score	0.88	0.93

- Accuracy: The ResNet-50 model outperformed the CNN model significantly, with an accuracy of 85.75% compared to 74%. This demonstrates that the ResNet-50 model is more effective in classifying emotions.
- Loss: The CNN model had a lower loss (0.68%) than the ResNet-50 model (1.72%). Lower loss signifies better performance, suggesting that the CNN model has better predictive capability.
- The validation AUC shows an overall upward trend, indicating that the model is improving its discriminatory power on unseen data between 0.6 to 0.7. Precision on the training set increases gradually.
- Time per step: Both models show relative performance in terms of time per step, with the ResNet-50 model slightly faster on average. However, the difference is negligible.
- Validation accuracy and loss: The ResNet-50 model demonstrates higher validation accuracy (85.8%) and lower validation loss (1.69%) compared to the CNN model (65% accuracy and 1.03% loss). This suggests that the ResNet-50 model generalizes better to unseen data [22-23]. The higher precision (0.94) of ResNet-50 denotes that the model is

more accurate in predicting the positive class without false alarms over CNN (0.89). ResNet-50 exhibits a superior recall compared to CNN, suggesting that it captures all relevant positive instances within the dataset more effectively.

- F1-score: This metric, being the harmonic mean of precision and recall, provides a balanced measure that is particularly useful when classes are imbalanced or when both false positives and false negatives need to be minimized. ResNet-50 demonstrated a higher F1-score relative to CNN, indicating that it maintains a better balance between precision and recall, making it a more reliable model overall.

The ResNet-50 model is the better choice for FER tasks, as it demonstrates improved feature extraction abilities and overall performance metrics compared to the CNN model.

IV. CONCLUSION AND FUTURE DIRECTIONS

The ResNet-50 model performs better in terms of accuracy and validation metrics, indicating excellent classification capability and better inference to unseen data. However, the CNN model showed better loss metrics, indicating better predictive capability in minimizing errors. Depending on the specific needs (e.g., prioritizing accuracy vs. minimizing loss), a researcher might choose between the two models. The ResNet-50 model emerges as the preferred solution for FER, leveraging its increased depth, feature extraction capabilities, and overall performance metrics. Future research could explore ways to further enhance the model's accuracy, potentially by incorporating existing ensemble learning techniques. Additionally, the focus should be given to the practical applications of these models by developing comprehensive datasets, leveraging transfer learning to train the models, and ultimately deploying them for real-time FER. Basic emotion recognition may lead to complex emotion recognition too.

REFERENCES

[1] P. Ekman, "Facial expression and emotion," *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993, https://doi.org/10.1037/0003-066X.48.4.384.

[2] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and Cognition*, vol. 17, no. 2, pp. 484–495, Jun. 2008, https://doi.org/10.1016/j.concog.2008.03.019.

[3] R. Plutchik, "A psycho evolutionary theory of emotions," *Social Science Information*, 1982.

[4] H. Zhang and M. Xu, "Modeling temporal information using discrete fourier transform for recognizing emotions in user-generated videos," in *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 629–633, https://doi.org/10.1109/icip.2016.7532433.

[5] D. Cheng, Y. Fan, S. Fang, M. Wang, and H. Liu, "ResNet-AE for Radar Signal Anomaly Detection," *Sensors*, vol. 22, no. 16, Jan. 2022, Art. no. 6249, https://doi.org/10.3390/s22166249.

[6] J. Miao, S. Xu, B. Zou, and Y. Qiao, "ResNet based on feature-inspired gating strategy," *Multimedia Tools and Applications*, vol. 81, no. 14, pp. 19283–19300, Jun. 2022, https://doi.org/10.1007/s11042-021-10802-6.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, Jun. 2009, https://doi.org/10.1109/cvpr.2009.5206848.

[8] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, Jun. 2014, pp. 1717–1724, https://doi.org/10.1109/cvpr.2014.222.

[9] J. Wu, Yinan Yu, Chang Huang, and Kai Yu, "Deep multiple instance learning for image classification and auto-annotation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3460–3469, https://doi.org/10.1109/cvpr.2015.7298968.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, https://doi.org/10.1109/cvpr.2016.90.

[11] J. Oheix, "Face expression recognition dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset.

[12] M. F. Aza, N. Suciati, and S. C. Hidayati, "Performance Study of Facial Expression Recognition Using Convolutional Neural Network," in *2020 6th International Conference on Science in Information Technology (ICSITech)*, Palu, Indonesia, Oct. 2020, pp. 121–126, https://doi.org/10.1109/icsitech49800.2020.9392070.

[13] Z. Y. Huang *et al.*, "A study on computer vision for facial emotion recognition," *Scientific Reports*, vol. 13, no. 1, May 2023, https://doi.org/10.1038/s41598-023-35446-4.

[14] D. O. Melinte and L. Vladareanu, "Facial Expressions Recognition for Human–Robot Interaction Using Deep Convolutional Neural Networks with Rectified Adam Optimizer," *Sensors*, vol. 20, no. 8, Apr. 2020, Art. no. 2393, https://doi.org/10.3390/s20082393.

[15] K. Mistry, L. Zhang, S. C. Neoh, C. P. Lim, and B. Fielding, "A Micro-GA Embedded PSO Feature Selection Approach to Intelligent Facial Emotion Recognition," *IEEE Transactions on Cybernetics*, vol. 47, no. 6, pp. 1496–1509, Jun. 2017, https://doi.org/10.1109/tcyb.2016.2549639.

[16] Y. Khaireddin and Z. Chen, "Facial Emotion Recognition: State of the Art Performance on FER2013." arXiv, May 08, 2021, https://doi.org/10.48550/arXiv.2105.03588.

[17] M. Dirik, "Optimized Anfis Model with Hybrid Metaheuristic Algorithms for Facial Emotion Recognition," *International Journal of Fuzzy Systems*, vol. 25, no. 2, pp. 485–496, Mar. 2023, https://doi.org/10.1007/s40815-022-01402-z.

[18] A. Alshammari and M. E. Alshammari, "Emotional Facial Expression Detection using YOLOv8," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16619–16623, Oct. 2024, https://doi.org/10.48084/etasr.8433.

[19] M. Talele, R. Jain, and S. Mapari, "Complex Face Emotion Recognition Using Computer Vision and Machine Learning," in *Harnessing Artificial Emotional Intelligence for Improved Human-Computer Interactions*, IGI Global, 2024, pp. 180–196.

[20] M. Tripathi, "Facial Emotion Recognition Using Convolutional Neural Network," *ICTACT Journal on Image and Video Processing*, vol. 12, no. 1, pp. 2531–2536, Aug. 2021, https://doi.org/10.21917/ijivp.2021.0359.

[21] S. Porcu, A. Floris, and L. Atzori, "Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems," *Electronics*, vol. 9, no. 11, Nov. 2020, Art. no. 1892, https://doi.org/10.3390/electronics9111892.

[22] M. a. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial Emotion Recognition Using Transfer Learning in the Deep CNN," *Electronics*, vol. 10, no. 9, Jan. 2021, Art. no. 1036, https://doi.org/10.3390/electronics10091036.

[23] C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[24] P. D. M. Fernandez, F. A. G. Pena, T. I. Ren, and A. Cunha, "FERAtt: Facial Expression Recognition With Attention Net," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, Jun. 2019, pp. 837–846, https://doi.org/10.1109/cvprw.2019.00112.

[25] L. V. L. Pinto *et al.*, "A Systematic Review of Facial Expression Detection Methods," *IEEE Access*, vol. 11, pp. 61881–61891, 2023, https://doi.org/10.1109/access.2023.3287090.

[26] S. W. Naidoo, N. Naicker, S. S. Patel, and P. Govender, "Computer Vision: The Effectiveness of Deep Learning for Emotion Detection in Marketing Campaigns," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, 2022, https://doi.org/10.14569/ijacsa.2022.01305100.

[27] A. Kaehler and G. Bradski, Learning OpenCV 3: Computer Vision in C++ with the OpenCV Library. O'Reilly Media, Inc., 2016.

[28] Madhivarman, "How to calculate the number of parameters in the CNN?," *Medium*, May 30, 2018. https://medium.com/@iamvarman/how-to-calculate-the-number-of-parameters-in-the-cnn-5bd55364d7ca.

[29] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, Jan. 1998, https://doi.org/10.1109/34.655647.

AUTHORS PROFILE

Milind Talele is a Ph.D. researcher in artificial intelligence with 20 years of experience in the IT industry in diverse domains and technologies. He has published several research papers on face emotion recognition and time series forecasting. He is proficient in project management and agile implementation across various industries. He holds a bachelor's degree in Electronics Engineering and an MBA in IT from Pune University.

Rajashree Jain currently works at the Symbiosis Institute of Computer Studies Research (SICSR), Symbiosis International University. Her research areas are microwave techniques, antennas, optimization, nature based algorithms, AI and sensor networks, IOT, data analysis, entrepreneurship and education. She has more than 25 research papers to her credit published in reputed journals and has presented research papers at various national and international conferences. She has served as chair or organizing member in various technical conferences and leadership summits. She is an active member of professional organizations such as IEEE.