# Real Time Facial Emotion Recognition using Deep Learning Models

Naveen N.C., PhD
Department of Computer Science and Engineering
JSS Academy of Technical Education, Bengaluru

Sai Smaran K.S.
Department of Computer Science and Engineering
JSS Academy of Technical Education, Bengaluru

Shamitha A.S.
Department of Computer Science and Engineering
JSS Academy of Technical Education, Bengaluru

## ABSTRACT
Facial emotion detection, a pivotal component of Artificial Intelligence (AI) and Computer Vision (CV), aims to recognize and identify human emotions from facial expressions. This paper presents an approach leveraging the Deep Learning (DL) models that includes Convolution Neural Network (CNN), Dual-Temporal Scale Convolutional Neural Networks (DTSCNN), Recurrent Neural Networks (RNN), and Residual Networks (ResNet-50) to achieve real-time and accurate emotion recognition. The primary objectives encompass real-time emotion recognition, high accuracy, low latency, and robustness to varied conditions. This paper performs experiments on benchmark datasets for evaluating each model considering accuracy, processing speed and facial orientations. This paper highlights the study of comparison between these models. The outcomes indicate that the CNN approach outperforms other methods, yielding superior accuracy and robustness. This research contributes in the advancement of facial emotion detection, with implications for applications in human-computer interaction, psychology, marketing, and healthcare. The CNN ensemble model represents a significant advancement in facial emotion detection, offering a comprehensive solution with broad implications across diverse domains. Its effectiveness highlights the significance of continuous exploration and refinement of deep-learning techniques to address complex tasks in CV and AI effectively.

## Keywords
Deep Learning, Emotion Detection, CNN, RNN, DTSCNN, ResNet-50

## 1. INTRODUCTION
Human expression is an essential facet of human communication, offering a detailed glimpse into individuals' thoughts, feelings, and intentions. Through facial expressions, individuals convey a myriad of information, complementing verbal communication and imbuing conversations with additional context, tone, and emotional depth. These expressions serve as invaluable indicators of emotional states, promoting better understanding and empathy in interpersonal interactions. Furthermore, they offer vital cues for assessing psychological well-being, assisting in the diagnosis of mental health conditions such as depression, anxiety, and trauma. In social dynamics, facial expressions play an integral role, influencing social perception, attraction, and rapport-building, thereby facilitating successful social interactions and fostering positive relationships and collaboration.

Facial emotion detection, emerging within the realms of CV and AI normally focus on the systems capable of accurately recognizing and interpreting human emotions through facial expressions. Leveraging facial recognition technology, these systems analyze various facial features such as expressions, movements, and patterns, translating them into evident emotional states that includes anger, sadness, fear, neutrality, disgust, surprise, and happiness. The real-time application of detecting facial emotion finds utility across diverse sectors including human-computer interaction, healthcare, education, and entertainment.

DL models and CNNs, have showcased remarkable progress in capturing intricate facial features and nuances, thereby leading to the progress of more robust emotion detection algorithms. This paper embarks on a comprehensive exploration of real-time application using DL approaches for detecting emotions, emphasizing on conducting a comparative study. The research focuses on underlying methodologies, challenges, and advancements within the field, emphasizing to illustrate the strengths and weaknesses of different DL approaches.

By leveraging the power of DL, this research endeavors to contribute to the refinement and enhancement of detecting emotions in real-time. Through a comparative analysis of DL models, the study seeks to identify the optimal approach for achieving heightened accuracy, reliability, and efficiency in facial emotion detection. Finally, the research endeavors to pave the way for advancements in recognizing emotions, facilitating more accurate predictions of individuals' mental states and advancing our comprehension of human emotion and behavior.

## 2. RELATED WORK
The field of Facial Emotion Recognition (FER) has witnessed remarkable advancements, driven by the convergence based on DL techniques, and sophisticated algorithms, opening the path for more precise and real-time emotion recognition systems. Each analysis has built upon the efforts of researchers who are pushing the threshold of knowledge. Amit Pandey et al. [1] were instrumental in highlighting challenges and opportunities in emotion detection. This paper also highlighted the variables that impact efficiency. The CNN algorithm was considered and it provided accuracy up to 55%. Following this, Dhwani Mehta et al.[2] discussed that using non-verbal cues like gestures and body movements also relies on robustness of an algorithm and sensitivity. Sensors are utilized to accurately capture emotions. Eventually, mixed reality devices like Microsoft HoloLens (MHL) were evolved in the field of Augmented Reality for observing emotions. Chirag Dalvi et al. [3] noted that there were limited studies providing a 360-degree overview of recognizing emotions. This paper covers a broad approach using various ML and DL techniques to perceive emotions for kids, adults, and senior citizens. Zi-Yu Huang et al. [4] conducted research on CV for perceiving the emotions. This study uses FER as the deep-neural network. The main aim of this analysis was to identify critical features and perform cross-database validation.

The maximum validation accuracy attained was 83.37%. The research conducted by Amr Mostafa et al. (referenced as [5]) addressed the issue of distinguishing between anger and disgust emotions. The main work carried out involved testing the emotions by combining visual features with RNN, that attains 82% accuracy. Following this, Sanchez-Ruiz et al. [6] highlight the active research in face expression FER technology, crucial for various applications requiring emotion verification. They propose a video-based FER system using temporal feature vectors from face landmarks, integrated into RNNs like Long Short Term Memory (LSTM) and Bidirectional Long Short Term Memory (BLSTM) for improved accuracy. Min Peng et al. [7] addressed facial micro-expression, which reveals the genuine emotions that people try to conceal. The DTSCNN was considered for this purpose, achieving 50.54% accuracy. Addressing problems such as poor generalization and low robustness, Bin Li and Dimas Lima [8] proposed a method using the Deep Residual Network Resnet-50. They utilized a new dataset comprising only 700 images to reduce training time, achieving 95.39% accuracy. Poonam Dhankhar [9] discussed the utilization of feature extraction of facial emotions using a blending of neural networks, namely Resnet-50 and VGG-16, which increased efficiency to 92.4%. This combination was considered as Resnet-50 alone provides an accuracy of 65.1%, and VGG-16 provides an accuracy of 59.2%. Ishika Agarwal et al. [10] adopted CNN to detect facial emotions in a large dataset consisting of 35k images. The system achieved an accuracy of 81.3% during the training phase and 69.2% during the validation phase, with a minimal loss of 0.35% during training and 1.25% during validation. Chahak Gautam and Seeja K R [11] proposed an approach to detect emotions by extracting features coupled with CNN. The techniques that are used to extract features are HOG and SIFT. The model, when tested, revealed its capability to recognize emotions with an accuracy of 98.48% and 91.43% using HOG-CNN for the CKplus and Jaffe datasets, respectively. With SIFT-CNN, accuracies of 97.96% and 82.85% were recorded for the CKplus and Jaffe datasets, respectively.

# 3. ABOUT THE DATASET

## 3.1 FER Dataset

The FER dataset represents a pivotal resource within the domain of facial emotion recognition, comprising an extensive collection of 35,887 images showcasing seven types of emotional states. These emotions encompass angry, happy, sad, fear, disgust, neutral, and surprise. Each image within the dataset is standardized to a dimension of 48x48 and converted to grayscale, a preprocessing step aimed at enhancing training efficiency and reducing processing overhead. The dataset is organized into several emotions as mentioned in table1.

**Table 1. FER Dataset**

| Emotions | Train | Test |
|---|---|---|
| Angry | 3,993 | 960 |
| Disgust | 436 | 111 |
| Fear | 4,103 | 1,018 |
| Happy | 7,164 | 1,825 |
| Neutral | 4,982 | 1,216 |
| Sad | 4,938 | 1,139 |
| Surprise | 3,205 | 797 |

The diverse distribution of images across different emotions led to varying levels of accuracy in detection; while some emotions were accurately detected, others exhibited lower accuracy rates. Specifically, the CNN algorithm achieved an overall accuracy of 78.06% when applied to this dataset, indicating both successes and challenges in facial expression recognition.

## 3.2 Skewed FER Dataset

In response to the challenges encountered with the original FER dataset, particularly in accurately predicting emotions like Disgust, measures were taken to enhance accuracy and expedite training speed. This involved performing dataset skewing. The resulting modified dataset is structured as in table 2.

**Table 2. Skewed FER Dataset**

| Emotions | Train | Test |
|---|---|---|
| Angry | 700 | 200 |
| Disgust | 436 | 111 |
| Fear | 700 | 200 |
| Happy | 700 | 200 |
| Neutral | 700 | 200 |
| Sad | 700 | 200 |
| Surprise | 700 | 200 |

The revised dataset demonstrates significant improvements, with all emotions now having an equal and substantial number of images suitable for training purposes. This dataset achieved a remarkable maximum accuracy of 96.03% when utilizing the CNN algorithm, showcasing a balanced and effective recognition of each emotion. The dataset is skewed based on relevancy, clarity and distinguishability of various images contained in the dataset. This outcome underscores the success of dataset skewing in addressing the initial challenges and enhancing the whole performance of facial emotion recognition.

# 4. METHODOLOGY

This research presents a relative study of the DL algorithms employed in FER. The research evaluates facial data using different models and interprets which model is the best fit. The methodology involves steps from dataset selection, data pre-processing, feature extraction, model selection, and model evaluation. The main seven emotions of a human being considered for evaluation are sad, happy, surprise, fear, angry, disgust, and neutrality.

The various DL models that are employed in this research are DTSCNN, RNN, ResNet, and CNN which are popular in image recognition applications due to their high accuracy. DTSCNN is a vibrant new field of research. DTSCNN that utilizes two convolutional network layers with distinct kernel sizes to capture short-term and long-term temporal dependencies simultaneously. It is particularly useful in tasks like speech recognition and audio processing, where understanding both local and global temporal patterns is crucial. By integrating multiple temporal scales, DTSCNN enhances the model's ability to extract meaningful features from sequential data. Its architecture typically has convolutional-layers, pooling-layers,

and optionally full connected-layers for the classification or regression tasks.

RNNs process sequential data by maintaining an internal state, capturing temporal dependencies. They have recurrent connections looping back to previous time steps, enabling memory retention. RNNs utilize a hidden state to store information from past time steps, allowing the parameter sharing across the sequence. Training involves Back Propagation Through Time (BPTT), propagating gradients backward through time steps. In image recognition, RNNs are often used in conjunction with CNNs for tasks like generating captions or descriptions for images and capturing temporal context in video classification.

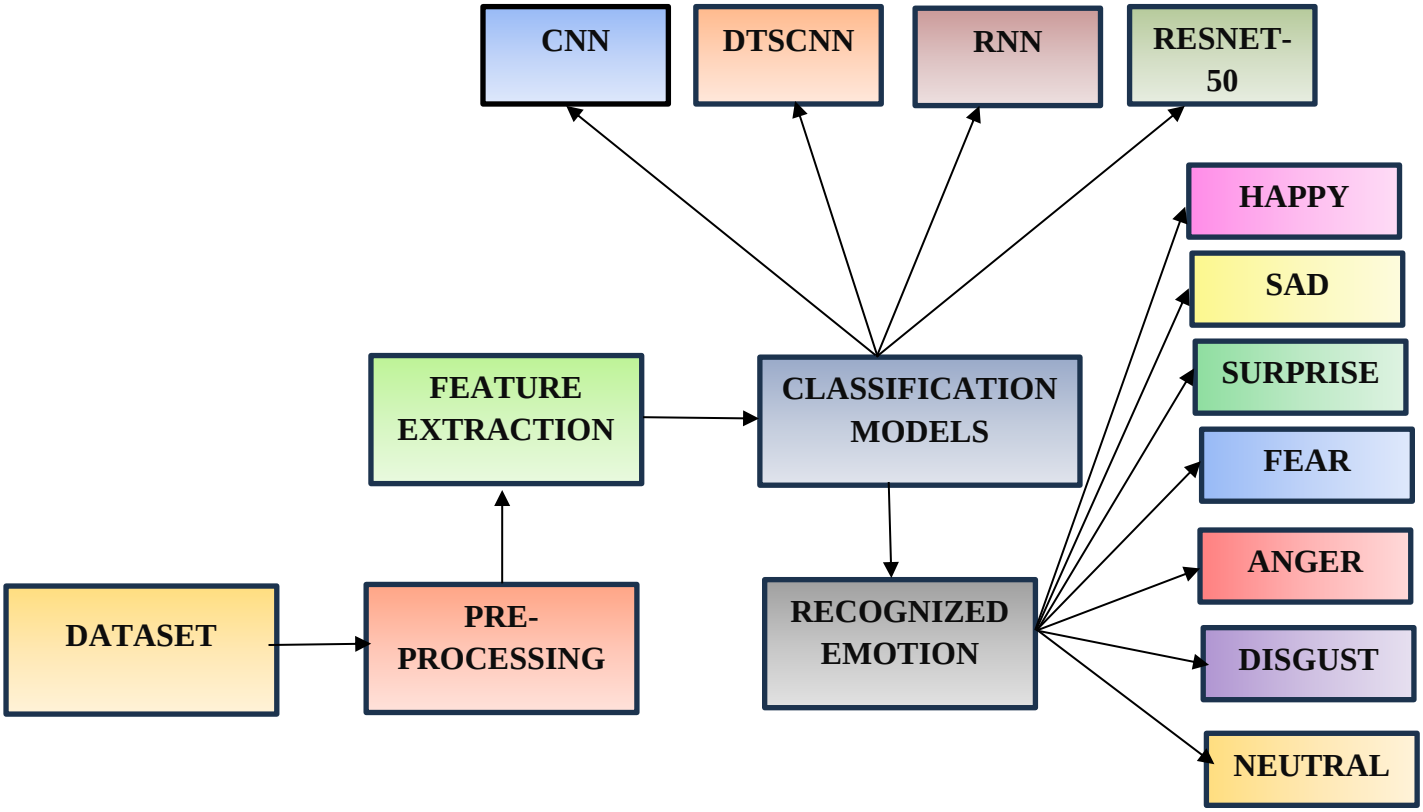


**Fig 1: Methodology**

ResNet-50 is a deep-convolutional neural-network architecture composed of 50 layers developed by Microsoft Research for image recognition tasks. It introduces residual connections, enabling training of very deep networks while mitigating the vanishing gradient problem, resulting in improved performance and accuracy on image classification tasks. ResNet-50 is renowned for its effectiveness in the tasks of image recognition. By incorporating residual connections, it enables training of deeper networks without encountering the vanishing gradient problem, leading to improved performance and accuracy.

CNN, a variant of Artificial Neural Network (ANN), comprises three main layers: convolutional layers, pooling-layers, and fully connected layers. The convolutional layer, the building block of CNN, applies learnable filters to input data, detecting features like edges and textures, producing feature maps. Stacking multiple layers creates hierarchical representations, capturing increasingly abstract features. Incorporating activation functions introduces non-linearity, enhancing the network's ability to learn complex patterns. Pooling layers in CNNs down sample feature maps, reducing spatial dimensions and computational complexity. The methods for pooling include max pooling, selecting the maximum-value within each region, and average pooling, computing the average. Pooling helps achieve translation invariance and reduce overfitting while preserving important features. Fully connected layers in CNNs connect every neuron in one layer to every neuron in the next layer, enabling high-level reasoning, crucial for classification tasks, mapping extracted features to output classes. CNN is mainly used in facial emotion recognition as it has the capability to process a huge amount of data with high accuracy.

## 5. RESULTS

| CNN | | DTSCNN | | RNN | | RESNET-50 | |
|---|---|---|---|---|---|---|---|
| Epochs | Accuracy | Epochs | Accuracy | Epochs | Accuracy | Epochs | Accuracy |
| 100 | 90.25 | 100 | 85.20 | 100 | 72.30 | 100 | 44.62 |
| 150 | 93.79 | 200 | 90.31 | 200 | 85.48 | 150 | 47.29 |
| 200 | 96.03 | 300 | 93.46 | 300 | 90.94 | 200 | 45.74 |

**Fig 2: Comparison study of Deep Learning Models**

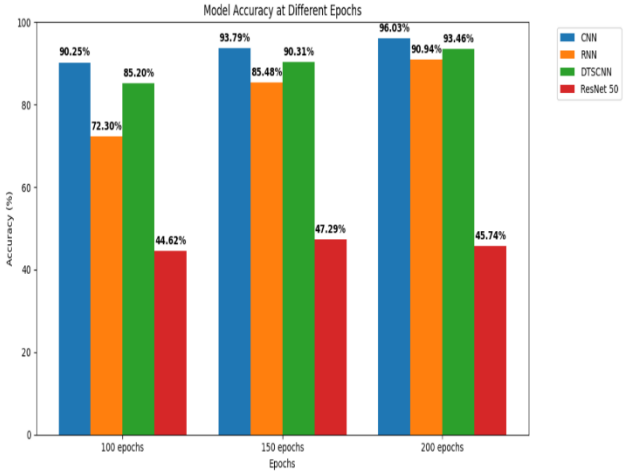


**Fig 3: Comparison study graph with epochs**

By comparing the above models, this research conveys that CNN recognizes facial emotions with the highest accuracy. CNN provides 96.03% accuracy for 200 epochs with high processing speed. DTSCNN provides 93.46% accuracy with good processing speed. DTSCNNs are still developing algorithms that require more research compared to CNN, which already dominates the image recognition technology. RNN provides 90.94% accuracy and this model has highest processing speed compared to other three models. RNNs is suitable to use alongside CNN for generating captions rather than acting independently, especially when there is temporal data. The processing speed of Resnet-50 is very low and has 45.74% accuracy. ResNet-50 takes more time and memory compared to CNN. Its deeper structure and larger number of parameters can lead to increased computational requirements during training and inference, making it less suitable for resource-constrained environments or real-time applications. Additionally, the deeper layers in ResNet-50 might capture more generic features rather than specific facial expressions, requiring extensive data and computational resources for effective training.

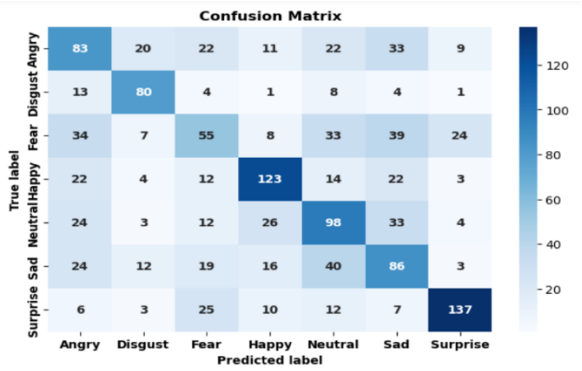


**Fig 4: Confusion matrix**

The confusion matrix in the Fig 4 shows that happiness, surprise, neutral, and sad emotions are predicted with high accuracy. Anger and disgust are predicted moderately well, whereas fear is predicted with less accuracy.

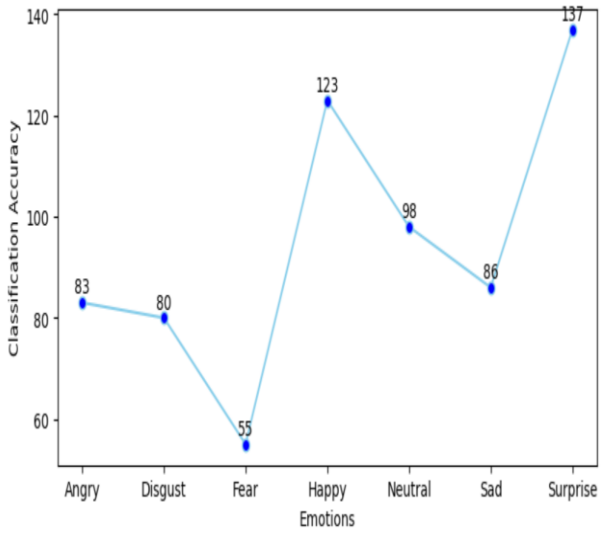


**Fig 5: Classification accuracy graph**

Its deeper structure and larger number of parameters can lead to increased computational requirements during training and inference, making it less suitable for resource-constrained environments or real-time applications. Additionally, the deeper layers in ResNet-50 might capture more generic features rather than specific facial expressions, requiring extensive data and computational resources for effective training. Hence, it can be concluded that CNN outperforms all other models and is most suited for facial emotion recognition.

## 6. CONCLUSION

This research work deals with the study comparison between the four models for facial-expression recognition: CNN, RNN, Resnet-50, and DTSCNN. The paper initially discussed the applications of facial-emotions, highlighting their significance in numerous fields such as human-computer interaction, and mental health assessment. Through rigorous experimentation and evaluation, the CNN model proves to be the efficient, consistently delivering accuracy rates exceeding 90% across various epoch configurations. However, challenges were encountered, particularly in handling variations in dataset composition and model complexity. For further research work different datasets maybe used for extensive evaluation of different DL techniques. Notably, the CNN model attained a peak accuracy of 96.03%, showcasing its reliability in perceiving facial-expressions. This study presents the critical role of model selection and the ongoing obstacles in attaining high accuracy for recognizing accurate emotions in real time. In future different ML methods as well as advanced DL algorithms may be considered and evaluated. In this research work 7 emotions were considered and research can be conducted with more number of emotions with a possibility of getting more accurate results.

## 7. ACKNOWLEDGEMENTS

The authors acknowledge the guidance provided by Dr.Abhilash C B, Associate Professor, Department of Computer Science and Engineering, JSS Academy of Technical Education Bengaluru, as well as the assistance of Mr.Rishav Kumar and Ms. Shilpa Gupta during the development of UI and analytics.

## 8. REFERENCES

[1] Pandey, Amit & Gupta, Aman & Shyam, Radhey. (2022). FACIAL EMOTION DETECTION AND RECOGNITION. 7. 176-179. 10.33564/IJEAST. 2022.v07i01.027.

[2] Mehta D, Siddiqui MFH, Javaid AY. Facial Emotion Recognition: A Survey and Real-World User Experiences in Mixed Reality. Sensors (Basel). 2018 Feb 1;18(2):416. doi: 10.3390/s18020416. PMID: 29389845; PMCID: PMC5856132.Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems

[3] C. Dalvi, M. Rathod, S. Patil, S. Gite and K. Kotecha, "A Survey of AI-Based Facial Emotion Recognition: Features, ML & DL Techniques, Age-Wise Datasets and Future Directions," in IEEE Access, vol. 9, pp. 165806-165840, 2021, doi: 10.1109/ACCESS.2021.3131733.

[4] Huang, ZY., Chiang, CC., Chen, JH. *et al.* A study on computer vision for facial emotion recognition. *Sci Rep* **13**, 8425 (2023). https://doi.org/10.1038/s41598-023-35446-4

[5] Badr, Amr & Khalil, Mahmoud & Abbas, Hazem. (2018). Emotion Recognition by Facial Features using Recurrent Neural-Networks.417-422.0.1109/ICCES.2018.8639182. Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[6] M. Sanchez-Ruiz, J. Flores-Monroy, M. Nakano-Miyatake, E. Escamilla-Hernandez and H. Perez-Meana, "Face Expression Recognition using Recurrent Neural Networks," 2023 46th International Conference on Telecommunications and Signal Processing (TSP), Prague, Czech Republic, 2023, pp. 148-153, doi: 10.1109/TSP59544.2023.10197740.

[7] Bin Li, Dimas Lima, Facial expression recognition via ResNet-50, International Journal of Cognitive Computing in Engineering, Volume 2, 2021, Pages 57-64,

[8] Dhankhar Poonam. "ResNet-50 and VGG-16 for recognizing Facial Emotions." (2019).

[9] I. Agrawal, A. Kumar, D. Swathi, V. Yashwanthi and R. Hegde, "Emotion Recognition from Facial Expression using CNN," 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC), Bangalore, India, 2021, pp. 01-06, doi: 10.1109/R10-HTC53172.2021.9641578.

[10] Poonam Dhankar, ResNet-50 and VGG-16 for recognizing Facial emotions, International Journal of Innovations in Engineering and Technology, Volume 13 Issue 4, Pages 126-130