Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.047326

ARTICLE

Faster Region Convolutional Neural Network (FRCNN) Based Facial Emotion Recognition

J. Sheril Angel[1], A. Diana Andrushia[1,*], T. Mary Neebha[1], Oussama Accouche[2], Louai Saker[2] and N. Anand[3]

[1]Department of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, 641114, India

[2]College of Engineering and Technology, American University of the Middle East, Egaila, 54200, Kuwait

[3]Department of Civil Engineering, Karunya Institute of Technology and Sciences, Coimbatore, 641114, India

*Corresponding Author: A. Diana Andrushia. Email: diana@karunya.edu

Received: 02 November 2023 Accepted: 18 March 2024 Published: 15 May 2024

ABSTRACT

Facial emotion recognition (FER) has become a focal point of research due to its widespread applications, ranging from human-computer interaction to affective computing. While traditional FER techniques have relied on handcrafted features and classification models trained on image or video datasets, recent strides in artificial intelligence and deep learning (DL) have ushered in more sophisticated approaches. The research aims to develop a FER system using a Faster Region Convolutional Neural Network (FRCNN) and design a specialized FRCNN architecture tailored for facial emotion recognition, leveraging its ability to capture spatial hierarchies within localized regions of facial features. The proposed work enhances the accuracy and efficiency of facial emotion recognition. The proposed work comprises two major key components: Inception V3-based feature extraction and FRCNN-based emotion categorization. Extensive experimentation on Kaggle datasets validates the effectiveness of the proposed strategy, showcasing the FRCNN approach's resilience and accuracy in identifying and categorizing facial expressions. The model's overall performance metrics are compelling, with an accuracy of 98.4%, precision of 97.2%, and recall of 96.31%. This work introduces a perceptive deep learning-based FER method, contributing to the evolving landscape of emotion recognition technologies. The high accuracy and resilience demonstrated by the FRCNN approach underscore its potential for real-world applications. This research advances the field of FER and presents a compelling case for the practicality and efficacy of deep learning models in automating the understanding of facial emotions.

KEYWORDS

Facial emotions; FRCNN; deep learning; emotion recognition; face; CNN

1 Introduction

Facial emotion recognition has become a fascinating area of research, with potential applications in user-aware marketing, health monitoring, and emotionally intelligent human-machine interactions. The ability of computer vision systems to recognize human emotions has garnered significant attention

from researchers. The ability to recognize emotions has several applications in human-computer interaction (HCI), including but not limited to affective computing, interactive video games, human-robot interaction, and medical diagnostics. Emotions are expressed in unimodal and multimodal ways, essential to many facets of human communication. Certain facial expressions convey universal meanings, and six primary universal emotions have been identified across all cultures: Anger, disgust, fear, happiness, sadness, and surprise [1].

Speech, gestures, and facial expressions are just a few examples of unimodal social behaviours in which emotions can be observed. Additionally, bimodal behaviours, such as speech accompanied by facial expressions, can successfully communicate emotions. Additionally, emotions can be communicated via various multimodal elements, such as audio, video, and physiological signs. Recent advances in artificial intelligence (AI) and deep learning (DL) models have vastly enhanced the understanding of the emotions displayed in facial images. Based on an ideal deep learning model, this study provides an insightful facial emotion recognition (FER) approach. Emotional states are important in the learning process. Positive emotions such as joy, acceptance, trust, and contentment can positively impact learning, but negative emotions can cause learning difficulties and disrupt the learning experience [2]. Recognizing and appropriately addressing emotions in educational environments is critical for improving learning results.

The primary motivation derives from emotions' profound influence on learning. The pleasant emotions promote improved learning outcomes, while negative emotions present obstacles. In educational environments, identifying and managing emotions is essential. A notable advancement lies in demonstrating the applicability of the developed FER system in educational environments. The system's ability to identify periods of annoyance, boredom, or bewilderment and provide focused interventions for improved engagement and knowledge retention addresses a critical need to enhance learning experiences through personalized support. The developed FER system holds immense potential in healthcare settings, particularly for real-time emotion monitoring during therapy sessions or medical consultations. It can assist healthcare providers in understanding patients' emotional states, providing valuable insights for personalized care and treatment plans [3].

The FER system directly impacts social robotics, where robots with advanced FER capabilities can better understand and respond to human emotions. This can revolutionize human-robot relationships, making interactions more intuitive and natural. Social robots could be deployed in various settings, including elderly care and companionship [4]. The ability to detect and interpret facial expressions has implications for public safety and surveillance. Integrating FER systems in surveillance cameras can assist in identifying potential security threats by analyzing the emotions of individuals in crowded or sensitive areas. For instance, autonomous driving systems use CNNs to detect and interpret road signs, obstacles, and other vehicles. Similarly, medical devices employ CNNs for tasks such as image classification in radiology, helping professionals detect abnormalities [5].

The value of an automated emotion detection system lies in its ability to identify emotions swiftly. Deep learning excels at feature extraction, yielding intricate image details. Face images were preprocessed, and features were retrieved using three distinct Convolutional Brain Organization models. AlexNet can be prepared, modified, or profound. Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Naive Bayes (NB), and Convolution Neural Network (CNN) were used to group these highlights. Morphological estimations were produced for benchmarking purposes using a standard morphometric methodology [6]. The model attained a comparable testing precision of 94.88%. Additionally, CNN models performed better than conventional morphometric assessments. The ANN classifier is found to work effectively with the recovered CNN

components. According to the preliminary findings, this computerized method for identifying facial emotions may be helpful.

The CNN-based face demeanor identification is performed in [7]. The results are evaluated using open-source databases such as CK+, MUG, and RAFD. The results indicate that the CNN approach is quite persuasive in image appearance recognizable proof on diverse public data sets, resulting in advances in look examination. To identify feelings from images, this work presented a deep learning engineering based on convolutional brain structures (CNN). The Facial Feeling Acknowledgement Challenge (FERC-2013) and the Japanese Female Facial Inclination (JAFFE) datasets are used to examine the demonstration of the suggested technique. The proposed model achieves correctness rates of 70.14 and 98.65 per cent, respectively, for the FERC-2013 and JAFFE datasets.

The deep learning method based on Convolutional Neural Networks (CNN) could augment emotion recognition using facial features. This dataset comprises roughly 32,298 images, each characterized by distinct aesthetic qualities. The image noises are minimized after using the correct preprocessing steps. The seven facial expressions, categorized by the Facial Action Coding System (FACS), were identified in this study without the streamlining technique. Still, the same seven feelings are revealed in their work [8].

Programmed emotion recognition based on appearance is an intriguing report developed and carried out in numerous sectors, including security, welfare, and human-machine collaborations. Experts in this field are interested in developing techniques to figure out and code and concentrating on looks to increase computer expectations further. Given deep learning's impressive track record, diverse models have been leveraged to amplify performance outcomes [8]. This work aims to research recent studies on automated FER using deep learning, underscoring design strategies, datasets, and outcomes. This study aims to help and guide researchers by assessing previous studies and sharing experiences to improve this field.

As researchers use deep learning techniques, there is a promise of discovering new avenues that could redefine the understanding and response to emotions across various domains. The recommended model achieves 88.43% recognition precision, surpassing the best-in-class techniques used for this study [9–11]. Notably, this methodology demonstrates robust performance in determining accurate facial articulations, even in conditions devoid of visible light. Because of the Quick RCNN procedure's ability to manage over-fitted preparatory data, it is possible to recognize impacted areas of the face even in the presence of noise, concealing, variety, size, and light variations [12]. To demonstrate the practicality of the introduced work, a careful investigation was performed compared to other techniques used to deal with classification strategies over a standard data set.

FER algorithms have traditionally relied on fundamental artificial intelligence approaches. On the other hand, the innovative, Faster Region-based Convolutional Neural Network (FRCNN) technique tries to recognize and distinguish facial expressions automatically. To detect faces, the model employs a Mask RCNN (Region-based Convolutional Neural Network) model, which is well-known for its accuracy in detecting facial regions. The FRCNN method outperformed classic handcrafted feature-based approaches regarding accuracy and robustness in recognizing and classifying face expressions. This demonstrates the utility of deep transfer learning in the FER problem. The study's findings add to the corpus of knowledge in facial expression detection and deep learning applications. This technique automatically recognizes emotions from facial images and opens up new possibilities for various sectors.

With AI's burgeoning growth, emotion recognition innovation has seen significant traction. As previous algorithms have progressively failed to match human demands, artificial intelligence and

deep learning calculations have witnessed enormous results in various applications like categorization frameworks, proposition frameworks, design acknowledgement, and pattern recognition. Emotions are essential in determining an individual's thoughts, activities, and attitudes. Using the benefits of deep learning, robust emotion detection systems can be designed by paving the way for myriad applications ranging from sentiment analysis to secure facial authentication with high precision.

The proposed study introduces a novel facial emotion recognition (FER) strategy based on Convolutional Neural Networks (CNNs), emphasizing the application of deep learning in understanding and interpreting facial expressions. Integrating Inception V3-based feature extraction represents an innovative approach to capturing intricate emotional features. The proposed model leverages its ability to capture spatial hierarchies within localized regions of facial features. The advantages of Inception V3 and faster region Convolutional Neural Networks (CNNs) are that they can be utilized to process extensive data and extract crucial facial information. The research evaluates the performance of the proposed Faster R-CNN-based methodology against traditional morphometric assessments and various machine learning classifiers, providing a comprehensive benchmark for the efficacy of deep learning in FER. The study surpasses conventional methods with an accuracy of 98.4%, showcasing the superior capabilities of the CNN-based approach.

2 Related Work

The FER system is implemented for various types of input datasets. In this section, the FER-based state-of-the-art methods are discussed. A deep CNN architecture for FER demonstrates its efficacy across several datasets [6]. The images were resized to 48 × 48 pixels after the facial landmarks from the data were extracted. The employed architecture consists of two layers for convolution pooling, two modules for inception styles, and convolutional layers of 1 × 1, 3 × 3, and 5 × 5. By locally applying convolution layers, they mitigated overfitting and optimized local performance.

Lopes et al. [7] looked into the effects of data preprocessing to enhance emotion classification before training the network. Before introducing the data to their CNN model, they incorporated a slew of preprocessing techniques, such as data augmentation, rotation correction, cropping, downsampling, and intensity normalization [7]. Their model comprises two convolution-pooling layers ending in two linked with 256 and 7 neurons. The best weight gained during the training phase is used during the test stage. This experience was evaluated using three readily available databases: CK+, JAFFE, and BU-3DFE. Researchers have demonstrated that combining these preprocessing techniques is more effective than doing it separately.

Yolcu et al. [11] proposed detecting the main face traits. They used three identical CNNs designated for specific facial features like eyebrows, eyes, or mouth. Before the images are uploaded to CNN, they are cropped, and key-point facial detection is performed. A second form of CNN was constructed to detect facial emotion, and the iconic face obtained in tandem with the raw image was employed. Combining raw and iconic face images proved superior results to standalone methods. The infra-red images are used by [12] to detect facial emotions using deep learning techniques.

Kim et al. [13] explored the fluctuation of facial expressions throughout emotional states and proposed a spatiotemporal architect based on a CNN-LSTM combination. CNN first learns the spatial aspects of the facial expression across all moving state frames, followed by an LSTM that preserves the complete sequence of these spatial attributes. Spatio-temporal convolutional with Nested LSTM (STC-NLSTM) [14] used a novel architecture based on three deep learning subnetworks:

3DCNN for spatiotemporal feature extraction, temporal T-LSTM for temporal dynamics preservation, and convolutional C-LSTM for modelling multi-level features. EEG signals-based deep learning framework used for the emotional classification of people with visual disabilities [15].

The researchers behind this study aimed to showcase the benefits of using an Extreme Learning Machine (ELM) classifier that utilizes fully connected layer computations from the pre-trained AlexNet-CNN (Convolutional Neural Network) model instead of the SoftMax layer of the Deep Convolutional Neural Network (DCNN). This experimentation was conducted within the Age-Invariant Facial Recognition (AIFR) context. The experimental results demonstrated that the ELM classifier combined with feature extraction using the pre-trained AlexNet-CNN model proved effective for facial recognition across age variations [16].

Xie et al. [9] investigated a unique approach for AIFR dubbed Implicit and Explicit Feature Refinement (IEFP). When facial features are removed from a face image, vital information about identity, age, and other traits is lost. In the context of AIFR, eliminating redundant data while retaining only the identifying data beneath the face feature is critical. The primary goal of this study was to show that the DCNN's SoftMax layer can be substituted with an ELM classifier that uses fully connected layer computations from the pre-trained AlexNet-CNN model. The DCNN is a widely used model in various computer vision tasks, including facial recognition. The SoftMax layer is typically employed as the final layer for classification in DCNN-based models.

Its numerous convolutional and fully linked layers make it appropriate for various image identification jobs. The researchers wanted to improve the accuracy and resilience of facial identification by using the F.C. layer calculations from the pre-trained AlexNet-CNN model for ELM classification. The experiment was carried out in the context of Age-Invariant face Recognition (AIFR), a problematic work due to the natural ageing process, which changes face appearances over time. The researchers gathered a broad dataset featuring facial images of people of all ages. They extracted deep features from these images using the pre-trained AlexNet-CNN model, obtaining high-level representations of facial characteristics.

However, the researchers sought to explore an alternative approach by incorporating ELM classification using the F.C. layer measurements from the pre-trained AlexNet-CNN model. The AlexNet-CNN model is a deep learning architecture that gained significant attention after its success in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2012. Its convolutional and fully connected layers make it appropriate for various image identification jobs. The researchers aimed to improve the accuracy and resilience of facial identification by using the F.C. layer calculations from the pre-trained AlexNet-CNN model for ELM classification. The experiment was carried out in the context of Age-Invariant face Recognition (AIFR), a problematic work due to the natural ageing process, which changes face appearances over time.

The researchers gathered a broad dataset featuring facial images of people of all ages [16]. They extracted deep features from these images using the pre-trained AlexNet-CNN model, obtaining high-level representations of facial characteristics. The ELM classifier was then employed to classify these deep features into specific age groups, effectively performing the facial recognition task across different ages. The experimentation results showcased the effectiveness of this approach, as the ELM classifier, combined with feature extraction using the pre-trained AlexNet-CNN model, yielded accurate and reliable results in age-invariant facial recognition.

The convolution neural network is predominately used in facial emotion recognition. It is leveraged in terms of cascaded networks and hybrid networks [17–20]. However, the limitations of these networks rely on lighting conditions and facial hair considerations. Many researchers have used

CNN for facial emotion recognition and applications [21–23]. From this lens, this study discusses a recent advancement in deep learning, precisely the FRCNN approach. The primary implementation transformed facial emotion identification. FRCNN can identify the pattern even if the input images are occluded. Understanding emotions through face images has enormous promise for improving human-computer interactions.

Recent progress in AI and deep learning has given rise to advanced methods for recognizing facial emotions, outperforming the old-fashioned practices that depended on manually crafted features and classifiers trained with image or video data sets. A Faster Region Convolutional Neural Network (FRCNN) architecture is used in the proposed FER system to capture spatial hierarchies within localised regions of facial features. Facial expression recognition is more accurate and efficient with this specific FRCNN architecture. The proposed FER system incorporates an Inception V3-based feature extraction method, which extracts relevant features from facial images for subsequent emotion categorization. The proposed strategy's efficacy is confirmed by extensive testing on Kaggle datasets. With an overall performance metrics of 98.4% accuracy, 97.2% precision, and 96.31% recall, the FRCNN technique shows excellent performance. This study presents an insightful deep learning approach to facial emotion recognition (FER), adding to the dynamic field of emotion detection technologies. The robustness and high precision shown by the FRCNN method emphasize its viability for practical use, pushing the boundaries of FER research and showcasing the effectiveness and applicability of deep learning models in automating the interpretation of facial emotions.

3 Methodology

A combination of pre-trained deep learning models (Inception V3 and Faster R-CNN) are used to classify facial emotions. The image dataset used in this study features a variety of images representing several emotion categories, including anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset is labelled, meaning each image corresponds to one of the emotion categories. However, there is an imbalance in the dataset, with some emotion categories containing much more images than others. Imbalanced datasets have the potential to degrade model performance and may necessitate careful handling during training and evaluation.

This research has developed a novel FRCNN technique for detecting and categorizing facial emotions. It consists of Adam optimizer to optimize hyper-parameters, Inception V3 for feature extraction, and FRCNN for classification. The following sections provide details about dataset distribution, pre-processing, Inception V3-based feature extraction, and Faster Region-based CNN. Fig. 1 shows the flow diagram of the proposed method.

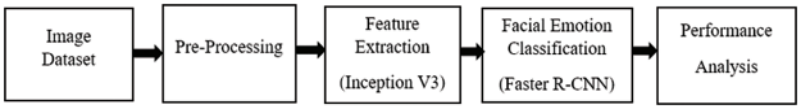

Figure 1: FRCNN-based facial emotion classification

3.1 Dataset Distribution

In this study, the suggested system was trained using a Kaggle dataset, freely available for scientific research upon request (https://www.kaggle.com). The dataset is separated for training, testing and validation sets. The dataset utilized in this study contains a variety of images representing various emotion categories: There are 4953 pictures for "Angry," 1460 images for "Disgust," 5344 images for

"Fear," 8448 images for "Happy," 4941 images for "Sad," 4945 images for "Surprise," and 5796 images for "Neutral." Table 1 shows the details of the dataset distribution for training, testing and validation. 80% of the dataset is used for training, 10% of the data is used for the testing phase, and 10% for validation.

Table 1: Details of dataset distribution

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Training set	3962	1168	4275	6758	3952	3956	4636
Testing set	498	147	536	846	497	505	589
Validation set	493	145	533	844	492	484	571

3.2 Preprocessing

The initial step is to ensure that all facial images are aligned and of the same size. The Viola-Jones technique detects the face, and the Hough transform is used to locate the centre of each eye for alignment. Facial image alignment and data preprocessing are required for facial recognition and analysis activities. The data is then prepared for analysis. This process includes cleaning the data, integrating disparate datasets, standardizing the data, and resizing the images as needed. Preprocessing data ensures it is clean and ready for analysis, decreasing errors and enhancing data quality. This resizing phase is required to ensure that the input images have a consistent and manageable size for the feature extraction and classification procedures that follow.

Additionally, the images are cleaned, including removing noise, artefacts, or irrelevant characteristics. Gaussian blur is strategically applied during preprocessing to mitigate pixel-level noise, enhancing dataset quality by convolving each pixel with a Gaussian kernel. It smooths fine-grained variations, reducing imperfections in facial images. Parameter tuning ensures optimal noise reduction while preserving essential facial features. This consistent application across the dataset maintains uniformity, promoting a cleaner and more robust dataset for improved facial emotion recognition model training and generalization.

3.3 Feature Extraction Using Inception V3

Inception V3 is a Convolutional Neural Network (CNN)-based deep learning model for picture categorization. It is an upgraded version of the original model Inception V1, which was introduced as GoogLeNet by a Google team in 2014. Inception V3 is a 48-layer deep pre-trained CNN model that was trained on the ImageNet dataset of 1,000,000 images. The third iteration of the Inception CNN model has been completed. The Inception V3 image recognition model's outstanding accuracy on the ImageNet dataset, which reaches 78.1%, is one of its primary achievements. This level of precision reflects the model's ability to recognize and classify objects in pictures.

The model implements a unique filter concatenation technique incorporating convolutional filters of 1×1, 3×3, and 5×5, followed by a pooling phase. Before applying 3×3 and 5×5 convolutions, 1×1 filters operate as a bottleneck, significantly lowering computing overhead. Asymmetric factorization process is adopted for the convolution operation in Inception V3

This design approach attempts to improve the model's efficiency and performance. Inception V3's design includes a sequence of channels comprising 1×1, 3×3, and 5×5 filters, followed by a final pooling layer. When 1×1 filters are used, bottleneck layers are introduced, drastically decreasing

processing costs while keeping feature representation. The amount of learnable variables are much reduced. The overall computation load is reduced because the kernel 5×5 is performed by smaller kernels. Data reduction is a common technique for shrinking the amount of a dataset so that it may be used for training and analysis. Another significant characteristic of the model is its ability to adjust the total number of pixels in an image. This is particularly beneficial when reducing image sizes or prioritizing computational efficiency [24–27].

Fig. 2 shows the architecture setup of Inception V3. The softmax regression is used to retrain in the final layer of Inception that produces probabilities based on the evidence gathered from the inputs. The evidence is determined by adding bias to a sum of weights identified by the intensity of pixels. The evidence is then processed using the softmax algorithm to get the final probabilities. Given a n-dimensional feature vector, the softmax function, or normalized exponential, yields a vector of the same dimension with values between 0 and 1.

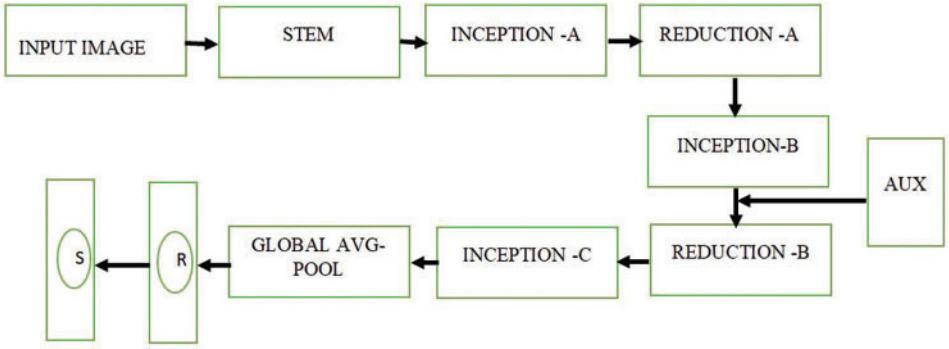

Figure 2: Architecture of Inception V3

3.4 Facial Emotion Classification Using Faster R-CNN

Faster R-CNN stands for "Region-based Convolutional Neural Network" and is a popular and prominent object detection framework in computer vision. The main contribution of Faster R-CNN was the introduction of a Region Proposal Network (RPN) that efficiently creates region proposals for potential objects in an image. This RPN is integrated with the following object detection pipeline, making the entire process more streamlined and faster than earlier approaches such as R-CNN and Faster R-CNN.

A Faster R-CNN framework often includes three major components. A deep convolutional neural network (CNN) that extracts significant information from an input image is referred to as a convolutional backbone. Faster R-CNN developed an end-to-end trainable architecture, meaning the entire model can be trained in a single optimization step. In contrast, prior approaches, such as R-CNN, required numerous phases of training and fine-tuning. In Faster R-CNN, the Region Proposal Network (RPN) dramatically enhanced the efficiency of region proposal creation. The RPN may create region suggestions significantly quicker than earlier methods by sharing convolutional features with the object identification module, resulting in real-time or near-real-time performance. Fig. 3 shows the overall architecture view of Faster R-CNN based facial emotion recognition.

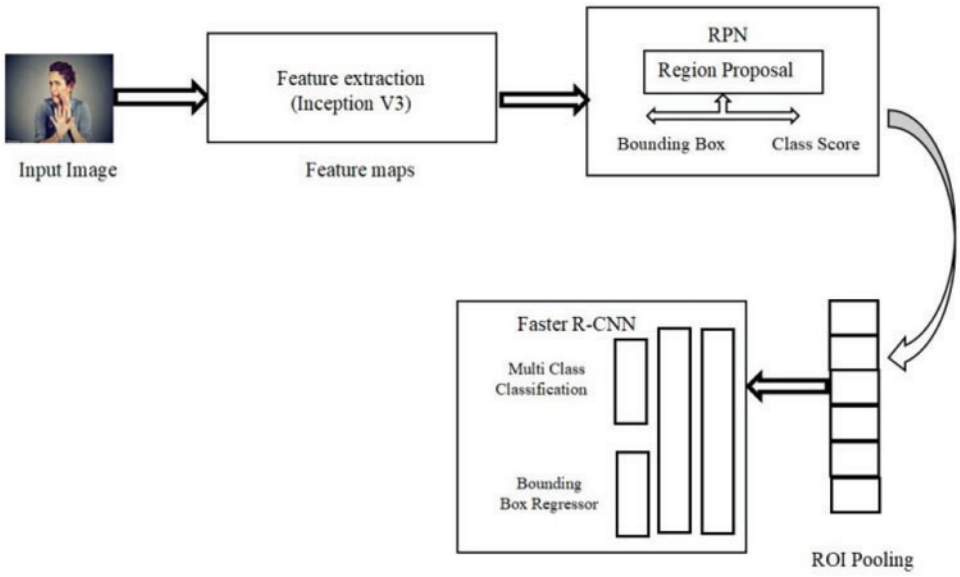

Figure 3: Architecture and process flow of Faster R-CNN-based facial emotion classification

Faster R-CNN is beneficial for object identification and can also be used for segmentation, which recognizes object instances and their matching pixel-level masks [28–30]. Faster R-CNN is more suitable for real-world applications due to its improved accuracy and speed, such as autonomous vehicles, surveillance systems, and object detection in videos.

Pseudo Code for Faster R-CNN:

Step 1: Assign the input image with several channels, height and width

Step 2: Extract the features using Inception V3

Step 3: Generate region proposals with bounding boxes

Step 4: Perform ROI-based pooling

Step 5: Classify the proposals

Step 6: Filter out the overlapping by applying non-maximum suppression

Step 7: Final detection outcome

The following are the critical steps of the RPN:

The RPN is utilized as an utterly shareable network that reduces the marginal costs. It creates object proposals which deal with a set of object classes against the background class. In order to improve the shape of the proposals, the overall pooling layers are tuned. The fully connected layers find the class score of the bounding box. So, the Faster R-CNN increase the accuracy of the detection tasks. Faster R-CNN is used to decrease the time of the training process, which interchanges the strategy of determining regions and running CNN. The convolutional sliding window is used to cover the entire feature map and to generate the region proposal. The characteristics of windows are used in the regression and classifier of fully connected layers. The output of the regressors concerning region coordinates (x_a, y_a, w_a, h_a) is given below:

$$O = \left[\frac{x - x_a}{W_{a,}}, \frac{y - y_a}{H_{a,}}, \frac{logW}{W_{a,}}, \frac{logH}{H_{a,}} \right] \tag{1}$$

$$O^* = \left[\frac{x^* - x_a}{W_a}, \frac{y^* - y_a}{H_a}, \frac{logW^*}{W_a}, \frac{logH^*}{H_a} \right] \tag{2}$$

where x_a, y_a, W_a, H_a are the anchor's centre coordinates, width, and height and x^*, y^*, W^*, H^* are the parameters of the ground truth bounding box.

The partial derivative of loss functions concerning input 'a' is given below:

$$\frac{\partial l}{\partial a_i} = \sum_x \sum_y (i = i^*(x, y)) \frac{\partial l}{\partial b_{xy}} \tag{3}$$

Let mini-batch ROI be x, and the output is b_y. The partial derivative is $\frac{\partial l}{\partial b_y}$. It is added if '1' is selected by the max pooling operation.

Integrating RPN into Faster R-CNN enhances the detection models' accuracy and speed. It generates region proposals for the prominent objects in the image. The RPN employs a predefined set of anchor boxes. Each anchor box has a set aspect ratio and scale that serves as a pattern for predicting potential object positions. These anchor boxes are densely packed across the feature maps' spatial dimensions. The RPN applies a tiny convolutional filter for each spatial point of the feature maps (usually 3×3). This sliding window method generates a set of fixed-size feature vectors for each anchor box. The model is trained for both region proposal and object detection. The facial emotion classification is performed through the stages of output block such as classifier and regressor.

3.4.1 Anchor Box Generation

Let A represent a collection of anchor boxes with varying scales and aspect ratios. The centre coordinates (x_a, y_a), width w_a and height h_a of each anchor box are specified.

The RPN predicts two variables for each anchor box: Box regression offsets (*dx, dy, dw, dh*) to adjust the dimensions of the anchor box and an objectness score that quantifies the chance of the anchor holding an object. Region Proposal: The RPN ranks and filters the anchor boxes based on the box regression offsets and objectness scores. Based on the objectness score, it chooses a set of top-scoring area ideas and uses non-maximum suppression (NMS) to eliminate extremely overlapped proposals.

The RPN creates the following for each anchor box: a.

Objectivity score: Object(a)-The likelihood that the anchor box contains an object.

Regression with bounding boxes: (a)-Changes to the anchor's coordinates to estimate the position of the item.

3.4.2 Pooling of Regions of Interest (ROI)

After getting region proposals from the RPN, the next step is to classify and improve the projected bounding boxes. Using ROI pooling, the feature maps from the convolutional backbone are used to obtain fixed-size feature vectors for each area proposal. This technique converts each region suggestion into a fixed-size feature representation that may be supplied into the succeeding item detection layers.

The ROI-pooled features are subsequently sent through fully linked layers that act as the object-detecting head. The object detection head has two functions: Box Classification: It uses softmax activation for multi-class classification to classify each region suggestion into different item categories

(e.g., person, automobile, cat). It refines the bounding box coordinates of each region proposal based on the expected box regression offsets from the RPN.

ROI pooling turns a ROI with feature map coordinates (x_r, y_r, w_r, h_r) into a fixed-size feature map *F_roi*. During training, the complete Faster R-CNN model is trained by an end-to-end training approach, which uses labelled images with ground-truth bounding boxes. The model is optimized using a combination of classification and regression loss functions. In essence, Faster R-CNN combines the efficiency of Region Proposal Networks (RPN) with the accuracy of object detection head components to enable real-time object recognition in pictures. It is a robust architecture that has served as the foundation for many future advances in object identification [31–35].

Faster R-CNN is a region-based object detection system. However, instead of detecting objects, it will be utilized to classify face expressions in this scenario. The Faster R-CNN model is fed the feature vectors extracted by Inception V3 for each facial image. For the specific purpose of facial expression categorization, the Faster R-CNN model has been fine-tuned or modified.

The Faster R-CNN model can simultaneously perform two tasks: Object identification (identifying area proposals) and facial emotion classification (classifying the emotion category for each region proposal). For each facial emotion category (angry, sad, disgusted, terrified, surprised, pleased, and neutral), the Faster R-CNN model will output region recommendations (bounding boxes) and the accompanying confidence values. The region proposals are generated using Inception V3 features and refined using the Faster R-CNN model.

3.4.3 Fully Connected Layer

Classification: The ultimately linked layers generate class scores *C_cls for each class*.

Bounding box regression: The fully connected layers predict changes to the bounding box of the ROI.

For each facial emotion category (angry, sad, disgusted, terrified, surprised, pleased, and neutral), the Faster R-CNN model will output region recommendations (bounding boxes) and the accompanying confidence values. The region proposals are generated using Inception V3 features and refined using the Faster R-CNN model.

4 Results and Discussions

The facial emotion recognition aimed to create a reliable and precise method of identifying human emotions in facial images. In affective computing and psychology research, Emotion recognition from facial expressions is essential in human-computer interaction. This study uses two cutting-edge deep learning models to complete the classification task: Inception V3 and Faster R-CNN. The experiment's dataset included various images representing the seven emotion classifications of anger, disgust, fear, happiness, sadness, surprise, and neutrality. The study concentrated on image pre-processing, feature extraction using Inception V3, and emotion classification using Faster R-CNN.

Fig. 4 shows the sample images in the dataset. It loads the original facial image into memory. Every image captures a face showcasing a distinct emotional expression. Grayscale conversion: The colour images are changed to grayscale for additional processing. Grayscale images contain one channel, making extracting features later on more straightforward while maintaining crucial facial characteristics. A thresholding technique is used to produce a binary image. Setting a threshold value transforms the grayscale image into a binary representation. When a value is above a threshold, a

pixel turns white (representing face characteristics), and when a value is below a threshold, a pixel turns black (representing background or noise).

Figure 4: Sample images of 7 facial emotions

Standardizing data is particularly crucial in facial emotion recognition (FER) to ensure consistent and reliable results across the diverse facial features associated with different emotional expressions. Standardizing data in facial emotion recognition is essential for creating models that consistently and reliably recognize emotions across various individuals, lighting conditions, and image variations. It promotes model robustness, efficiency, and interpretability, contributing to the overall effectiveness of facial emotion recognition systems. Facial emotion recognition models are sensitive to variations in lighting conditions, facial orientations, and other factors. Standardizing data helps reduce the impact of these variations, promoting the robustness of the model. This is crucial for consistent performance in real-world scenarios where facial expressions may differ in lighting and pose.

4.1 Training Phase

Entire dataset is divided for training phase, testing phase and validation phase. 80% of the dataset is used for training, 10% of the data is used for the testing phase, and 10% for validation. In order to get the accurate results, four fold cross validation is adopted. Henceforth all the images of dataset are included in the training phases. The Inception V3 model is leveraged as feature extraction process. It consists of multiple convolutional and pooling layers, followed by fully connected layers. The model implements a unique filter concatenation technique incorporating convolutional filters of 1×1, 3×3, and 5×5, followed by a pooling phase. The input image with size $256 \times 256 \times 3$ is given to Inception V3 network. The network consists of symmetric and asymmetric blocks which includes convolution stem layers, inception A block, reduction A block, inception B block, reduction B block and inception C block.

The global average pooling is followed by these blocks. $1000 \times 1 \times 1$ is the size of the fully connected layer. The regularization, factorization and dimension reduction techniques are adopted. Asymmetric factorization process is adopted for the convolution operation in Inception V3. To enhance data throughput within the network, the inception model introduces a layer that is both depth wise and width wise, thereby constructing a network that effectively manages the balance between depth and width for optimized performance. The pre-trained parameters of Inception V3 is used and immobilize the layers to prohibit their modification throughout the training process.

The extracted features map from Inception v3 is given to RPN of FRCNN. The region proposal network generate bounding boxes for potential facial regions of interest. ROI pooling extract the features from the potential facial regions by using the extracted features from Inception v3 model. The RPN tests feature maps with bounding boxes of various scales to recommend the ROIs. The CNN process the ROI features and process further by using fully connected layers. The tuning parameters of Inception V3 and FRCNN is highlighted in Tables 2a and 2b. Initially the model is trained for 10 epochs and the experimental results are analyzed. The overall performance of model is analyzed for 30 epochs.

Table 2: Tuning parameters of proposed model

(a) Inception V3	
Tuning parameters	Values
Batch size	32
Learning rate	0.001
Drop out	0.2–0.5
Regularization technique	L2
Number of epochs	30
% of training and testing samples	80% and 20%
(b) Faster R-CNN	
Tuning parameters	Values
Batch size	128
Learning rate	0.001
Weight decay	0.0001
Anchor sizes and ratios	[32, 64, 128] & [1:1, 1:2, 2:1]
ROI threshold	0.5
Number of epochs	30
% of training and testing samples	80% and 20%

4.2 Testing Phase

The major metrics of the deep learning model are True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) for each emotion category based on the model's predictions, and the ground truth labels are calculated. Precision (P) gauges how accurately the good forecasts came true. The proportion of true positive cases that are accurately detected is measured by recall (R), also known as sensitivity or true positive rate:

The F1-score, the harmonic mean of recall and precision, is used to balance these two metrics: The F1-score is derived by multiplying the precision and recall scores by two.

The facial emotion categorization model underwent training over 30 epochs. During the training phase, an epoch is a single pass of the entire dataset through the neural network. To track the model's training progress and choose the right number of epochs to get the best accuracy and convergence, the model's performance was evaluated after each epoch. The training and validation losses were recorded after each epoch to track the model's convergence. The validation loss measures the error on the hidden validation data. In contrast, the training loss measures the difference between the model's predictions and the ground truth labels on the training data. As training continued, both losses declined, demonstrating effective learning and generalization by the model.

A confusion matrix is commonly used for binary classification problems with two classes: Positive (P) and negative (N). It can also be extended to challenges involving more than two classes in multi-class categorization. To guarantee that the model in facial emotion recognition (FER) learns to recognize emotions across all classes, it is imperative to address imbalanced data. A helpful method for

overcoming the difficulties caused by unbalanced datasets is the cross-validation approach, which is applied during model training and evaluation. The confusion matrix shows the detailed performance of the proposed model. The model effectiveness is checked by the performance metrics of accuracy, precision, recall and F1-score. The individual class confusion matrix for seven facial emotions are shown in Fig. 5. The high values for these metrics indicate that the classifier is proficient at correctly classifying both positive and negative instances, with minimal instances of incorrect predictions. Consequently, the evidence from the confusion matrix and the associated performance metrics lead to the conclusion that the classifier's performance is robust.

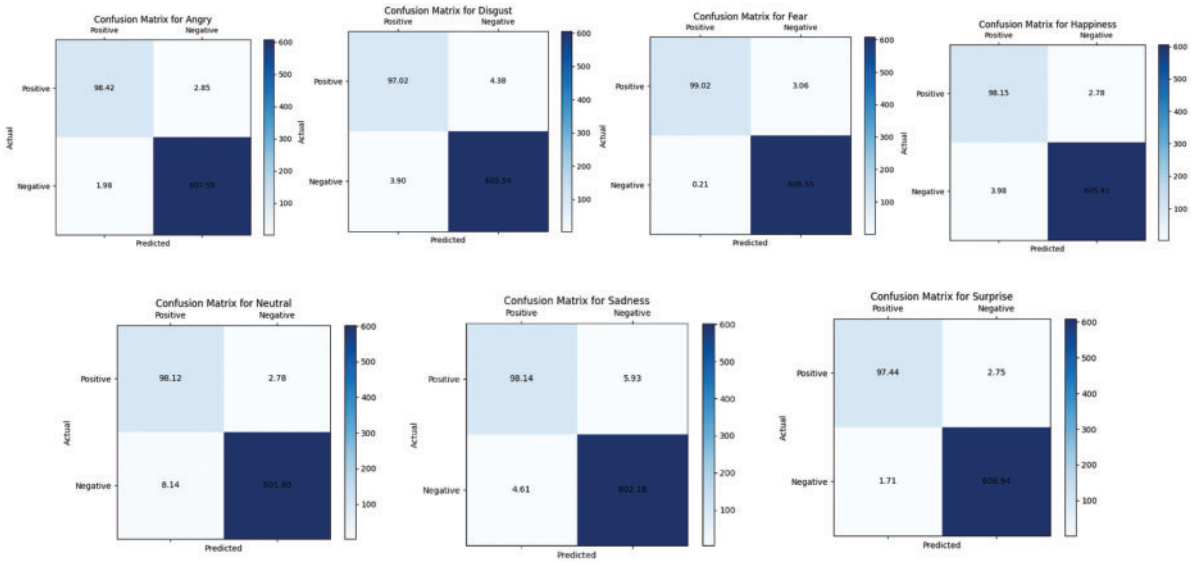

Figure 5: Confusion matrix of seven classes of facial emotions

Four-fold cross-validation is used to evaluate the performance of the facial emotion system. The total dataset is separated into four subsets, which are created randomly. From the four subsets, three subsets are used for the learning phase, and others are used for the testing phase. It is repeated four times. From the four subsets, each subset is used as a validation set for one time. Each fold sustains the class distribution and prevents the overfitting of each class. The seven distinct emotion classes are classified accurately. After adopting four-fold cross-validation, the images are not overlapped between different classes.

The performance metrics of proposed model for each epochs are given in Table 3. The epoch wise accuracy is shown. It ensures that the number or epochs increases, the increasing accuracy rate in training and validation phase. At the initial stages of training, the model's accuracy on both the training and validation sets is relatively low, with the first epoch yielding 41.34% and 42.43%, respectively. As the epochs progress, there is a clear trend of improvement. By the 5th epoch, the training accuracy slightly surpasses the validation accuracy, which could be an early sign of the model beginning to overfit to the training data. However, the validation accuracy continues to improve, reaching 58.45%, which suggests that the model is still generalizing well to unseen data at this point.

Table 3: Epoch-wise accuracy

Epoch	Training accuracy	Validation accuracy
1	41.34	42.43
2	43.65	45.48
3	46.67	49.88
4	50.96	54.66
5	51.29	58.45
6	58.29	59.11
7	63.77	61.43
	.....	...
20	91.45	82.41
	.....	.....
30	98.41	90.76

As the training continues towards the 20th epoch, the training accuracy reaches a high of 91.45%, while the validation accuracy is also robust at 82.41%. This indicates that the model has learned to predict the training data with high accuracy and is also performing well on the validation set, which is crucial for ensuring that the model will perform well on new, unseen data. By the 30th epoch, the training accuracy is at an impressive 98.41%, and the validation accuracy is at 90.76%. The consistent increase in validation accuracy suggests that the model is generalizing well and not merely memorizing the training data.

The proposed model flawlessly fits the input data. The performance metrics are obtained after performing the cross-validation steps. As the number of classes is high, the four-fold cross-validation is adopted. The proposed method achieved 98.4% classification accuracy in the minimum number of epochs. Fig. 6 shows the confusion matrix of the proposed work after performing four-fold cross-validation. The bold values represent significant values of the classes. Based on the comparative analysis, the proposed work is highly fit for real-time scenarios. Table 4 shows the performance metrics of four-fold: Precision, recall, f-measure, and accuracy.

Backpropagation with a conjugate gradient approach was used to train FRCNN. The categorization results are subject-dependent. Emotion recognition based on facial expressions across all photographs is far more helpful. The proposed model produces the accuracy of 98.4%, precision of 97.2%, and recall of 96.31%. The results are excellent, particularly for the FRCNN classifier. In such situations, calculating the confusion matrices becomes essential to determine which emotions are most easily identified and which are more difficult to distinguish.

The Accuracy graph indicates that maximum accuracy for the training dataset is achieved at epoch 25, while the utmost accuracy for validation is achieved at epoch 30. Notably, the training accuracy for the proposed method is higher than the validation accuracy. Fig. 7 highlights the accuracy and loss graphs of the training and validation phases. The Model loss graph shows that the training loss of the proposed method drops to the minimum value when the epochs increase. The training loss is 0.231. It indicates that the training loss is significantly lower than the validation loss.

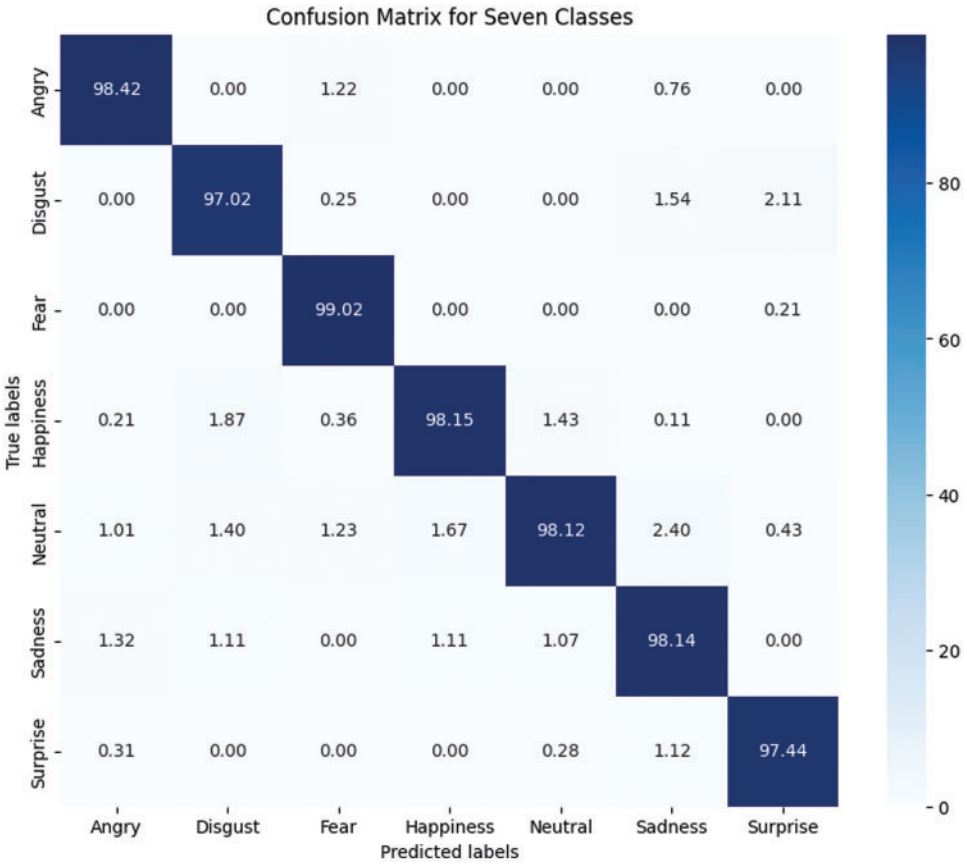

Figure 6: Confusion matrix of proposed facial emotion recognition system

Table 4: Four-fold cross-validation results

Folds	P%	R%	F1%	A%
Fold1	98.12	97.32	98.01	98.65
Fold2	93.67	97.45	96.76	97.92
Fold3	98.91	98.11	98.45	98.61
Fold4	94.54	95.23	95.45	97.83

4.3 Comparative Analysis

The proposed Faster R-CNN-based facial emotion recognition method is classified into seven types of emotions. Out of seven classes, two of the images are misclassified. In the class, one of the images is misclassified as a surprise. The widened eye is the critical feature of surprise. This feature appears common in both courses with slight variation. In the class of sad, an image is misclassified as disgust. The eyes that appear in the original image are too shrunken, so the emotion is identified as disgust. In addition, the mouth is over-embellished. Hence, the proposed technique is classified as disgust. The images in the remaining classes are correctly identified by their facial expressions. The results of the proposed work is compared with eight state-of-the-art methods. All the comparison

methods are taken concerning the CNN architecture and its modifications. Table 5 summarizes the results of the comparison method with the adopted methodology and performance metrics.

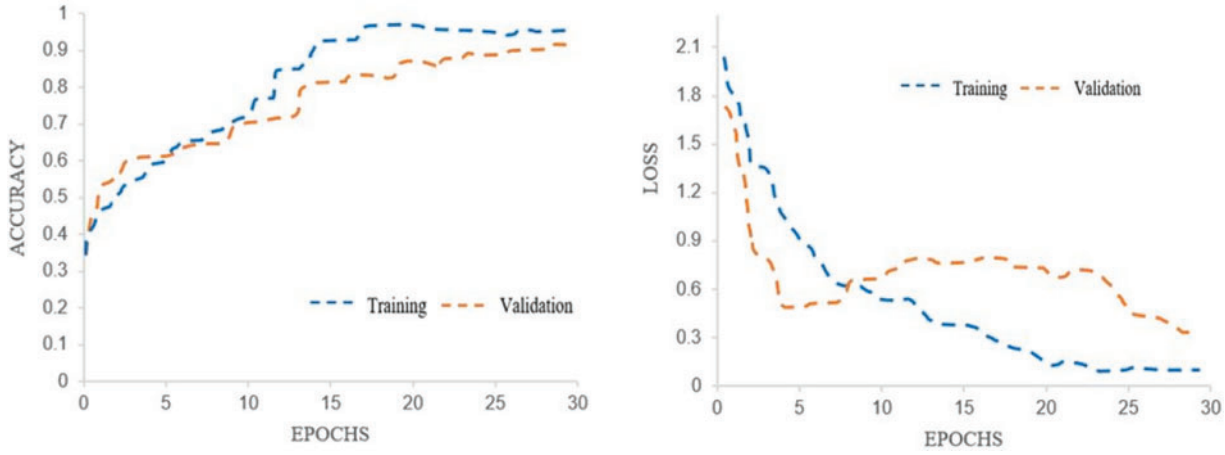

Figure 7: Accuracy and loss graphs of the proposed model

Table 5: Comparative analysis of proposed method with state-of-the-art methods

Sl. no.	State-of-the-art methods	Methodology	Accuracy
1	Kong (2019) [36]	LBP + CNN	91.28%
2	Ruiz-Garcia et al. (2017) [37]	Stacked convolutional auto-encoder (SCAE)	92.52%
3	Li et al. (2018) [38]	Attention-based convolutional neural network (ACNN)	91.64%
4	Jain et al. (2018) [39]	Hybrid CNN + RNN	94.91%
5	Cai et al. (2018) [40]	SBN-CNN	95.24%
6	Yolcu et al. (2019) [11]	Three CNN	94.44%
7	Uddin et al. (2017) [41]	Deep belief network (DBN)	96.25%
8	Jain et al. (2020) [42]	LSTM-CNN	96.17%
9	**Proposed**	Inception V3 + Faster R-CNN	98.4%

The proposed method is compared with eight state-of-the-art facial emotion recognition techniques. Many researchers worked in facial emotion recognition using machine learning and deep learning methodologies. CNN is used by many researchers with certain modifications. Kong [36] used a Local Binary Pattern (LBP) for the feature extractor. It is added in the fully connected layer of CNN for detecting facial expressions. Due to the integration of improved LBP with CNN, the overall method produced 91.28% accuracy in the facial expression detection task. The weights of CNN are initialized with the help of stacked convolutional auto-encoder weights [37]—these types of modifications produced better results than the general CNN of random weights. Li et al. [38] integrated attention mechanism-based focus function with CNN. The occluded regions in the face were considered to detect the facial expressions effectively. The areas of interest concerning the faces

are included by considering patch-based ROIs and global-local-based ROIs. It produced a detection accuracy of 91.64%.

The sparse batch normalization (SBN) based CNN is proposed by Cai et al. [40]. This network starts with two successive convolution layers, followed by a max pooling layer. To tackle overfitting, dropout is implemented in the middle of three fully connected layers. The SBN-CNN method achieved a detection accuracy of 95.24% for facial expression detection. Three CNN architectures are combined to detect the facial expression Three CNN architectures are combined to detect the facial expression in Yolcu et al. [11] method. The initial CNN identifies the critical point of the face. The eyebrows, eyes, and mouth are analyzed separately by three CNNs. The iconic face is obtained after combining the results of three CNNs. The detection accuracy of the method is 94.4%.

Fig. 8 shows the comparison chart of the proposed method and state-of-the-art methods. The directional patterns of each pixel are obtained and processed through discriminant analysis. After extracting the features, the feature set is given to a deep belief network (DBN) for learning facial emotions [41]. At the final stage, the DBN classified the facial emotions with an accuracy rate of 96.25%. Jain et al. [42] utilized a multi-angle optimal deep learning approach to detect facial expressions. LSTM-CNN model to classify the facial emotions and achieved an accuracy of 96.17%. The hybrid architecture of convolution neural network (CNN) and recurrent neural network (RNN) was used for FER in Jain et al.'s work. This hybrid architecture produces better results in comparison with non-hybrid methods.

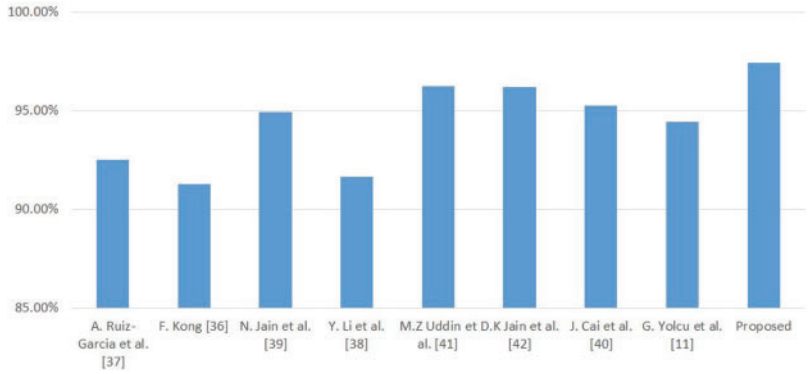

Figure 8: Performance metric comparison

4.4 Discussions

The recognition of facial emotions is an integral part of human-computer interaction, with applications ranging from virtual assistants to mental health monitoring. This work provides an in-depth examination of a novel technique for facial emotion recognition that employs the Inception V3 deep learning architecture in conjunction with Faster R-CNN object detection. The suggested model achieves 98.4% accuracy, indicating its potential for real-world applications. Compared to several standards and deep learning-based methods, the proposed FER method using Inception V3 and Faster R-CNN displays excellent accuracy and resilience. It can determine emotions by presenting the results in confusion matrices. The proposed work develops a FER system using a Faster Region Convolutional Neural Network (FRCNN) and design a specialized FRCNN architecture tailored for facial emotion recognition, leveraging its ability to capture spatial hierarchies within localized regions of facial features.

Emotions such as sadness and anxiety proved to be the most challenging to recognize accurately. Neutral and surprise emotions frequently misclassified them. Additional tests were conducted to investigate the feasibility of discriminating between specific emotions. The classifier was tested in paired combinations to pinpoint the sources of misclassification. The results from the FRCNN classifier confirm that most errors occur between two pairs: Sadness-neutral and surprise-fear. The expressions of surprise and terror are remarkably similar, with an open mouth and a raised eyebrow. Analogous alterations in the same facial expression significantly impact classification accuracy. In sadness and neutral feelings, inadequate changes can cause accuracy to deteriorate. The coefficient that should best distinguish between neutral and sad expressions. The results for some subjects were notably less accurate than for others, especially concerning characteristics like raised eyebrows. External factors such as the user's facial hair or skin colour may also influence the quality of emotion classification. The extensive study and detailed examination of the model's performance significantly contribute to advancing the field of emotion recognition and open the door to practical applications.

4.5 Limitations and Future Directions

The proposed method uses Inception V3 as a feature extractor and Faster R-CNN as a classifier. Even though the proposed model produces better classification accuracy, few limitations exist. The proposed model can effectively classify the seven facial expressions (anger, disgust, fear, happiness, sadness, surprise, and neutral). This model may struggle to organize the unseen facial expressions. The dataset can be enriched by adding more expression images regarding demographics, environmental conditions and expressions. It will enhance the proposed model's real-time deployment with better generalization and outcome.

There are several fascinating possible enhancements and future approaches for the facial emotion categorization study using Inception V3 and Faster R-CNN. Here are some potential future horizons:

More extensive and diverse dataset: The model's performance can be significantly improved by expanding the dataset with a more comprehensive and varied collection of facial images. A more extensive dataset would record various facial expressions and increase the model's capacity to extrapolate across multiple people and ethnicities. The mixed emotions of each class can be added. Mixed emotions include sorrow with surprise, shock with pleasure, etc.

Real-Time Implementation: Creating a real-time facial emotion recognition system that can examine facial expressions in real-world situations or live video streams has direct applications in virtual reality, emotion-aware technology, and human-computer interaction. Intensity Prediction of Emotions: Beyond categorizing emotions into distinct groups, anticipating the intensity or strength of emotions could offer more granular insights into people's emotional states. By exploring these directions, this work can be elevated to deliver emotion recognition technologies that are both versatile and excel in diverse real-world contexts.

5 Conclusion

FER is one of the vital research areas in the field of human-computer interaction. The proposed study introduces a customized FRCNN focused on the facial region for emotion identification. They are designing a specialized FRCNN architecture tailored for facial emotion recognition, leveraging its ability to capture spatial hierarchies within localized areas of facial features. The model was trained, tested, and validated on an image dataset, successfully distinguishing seven distinct facial expressions. The model can fit the data and generalize to additional information with identical training and validation accuracy. With the integration of the Adam optimizer, the model drastically reduced its

loss, achieving an impressive 98.4% accuracy in facial emotion recognition. The importance of this work extends beyond image-based emotion identification. The presented approach has the potential to be developed and applied to video sequences, allowing for real-time facial emotion recognition. Such capabilities can open up many options for various applications, including emotion analysis in video streams and real-time input assessment. For example, it could help healthcare provider's measure patients' emotional states during therapy sessions or medical consultations, delivering significant information.

Furthermore, this facial emotion recognition technique could be helpful in social robotics and artificial intelligence. Robots and virtual agents with this technology could better understand and respond to human emotions, leading to more empathetic and natural interactions. Such advancements can revolutionize human-robot relationships, making them more intuitive and comfortable for users. Its accuracy and generalizability make it an essential tool in many industries, including optimizing user experience, healthcare, and social robots. Recognizing emotions in real-time within video sequences presents intriguing possibilities for practical deployment across diverse fields. Nevertheless, it is imperative to approach its use cautiously, ensuring data privacy and the ethical employment of the technology.

Acknowledgement: The authors would like to express their gratitude for the valuable feedback and suggestions provided by all the anonymous reviewers and the editorial team.

Funding Statement: This research received no external funding.

Author Contributions: Conceptualization, methodology and validation, J. Sheril Angel, A. Diana Andrushia; resources, data curation, and visualization, J. Sheril Angel, A. Diana Andrushia. T. Mary Neebha; writing—original draft preparation, J. Sheril Angel; writing—review and editing, formal analysis, A. Diana Andrushia. T. Mary Neebha, N. Anand, Oussama Accouche, Louai Saker. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The training dataset can be accessed from https://www.kaggle.com.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. R. Khan, "Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges," *Information*, vol. 13, no. 6, pp. 268, 2022. doi: 10.3390/info13060268.

[2] N. Roopa, "Emotion recognition from facial expression using deep learning," *Int. J. Eng. Adv. Tech.*, vol. 8, no. 6S, pp. 91–95, 2019. doi: 10.35940/ijeat.f1019.0886s19.

[3] P. Adibi *et al.*, "Emotion recognition support system: Where physicians and psychiatrists meet linguists and data engineers," *World J. Psychiatry*, vol. 13, no. 1, pp. 1–14, 2023. doi: 10.5498/wjp.v13.i1.1.

[4] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Front. Rob. AI*, vol. 145, pp. 241, 2020. doi: 10.3389/frobt.2020.532279.

[5] T. K. Arora *et al.*, "Optimal facial feature based emotional recognition using deep learning algorithm," *Comput. Intell. Neurosci.*, vol. 20, no. 2, pp. 1–10, 2022. doi: 10.1155/2022/8379202.

[6] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter Conf. App. Comput. Vision (WACV)*, Lake Placid, NY, USA, 2016, pp. 1–10. doi: 10.1109/WACV.2016.7477450.

[7] 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, no. 12, pp. 610–628, 2017. doi: 10.1016/j.patcog.2016.07.026.

[8] A. Fathallah, L. Abdi, and A. Douik, "Facial expression recognition via deep learning," in *2017 IEEE/ACS 14th Int. Conf. Comput. Syst. App. (AICCSA)*, Hammamet, Tunisia, 2017, pp. 745–750.

[9] J. C. Xie, C. M. Pun, and K. M. Lam, "Implicit and explicit feature purification for age-invariant facial representation learning," *IEEE Trans. Inf. Foren. Secur.*, vol. 17, pp. 399–412, 2022. doi: 10.1109/TIFS.2022.3142998.

[10] S. A. Hussain and A. S. A. Al Balushi, "A real-time face emotion classification and recognition using deep learning model," *J. Phys.: Conf. Series,*, vol. 1432, no. 1, pp. 012087, 2020.

[11] G. Yolcu *et al.*, "Facial expression recognition for monitoring neurological disorders based on convolutional neural network," *Multimed. Tools Appl.*, vol. 78, no. 22, pp. 31581–31603, 2019. doi: 10.1007/s11042-019-07959-6.

[12] A. Bhattacharyya, S. Chatterjee, S. Sen, A. Sinitca, D. Kaplun and R. Sarkar, "A deep learning model for classifying human facial expressions from infrared thermal images," *Sci. Rep.*, vol. 11, no. 1, pp. 20696, 2021. doi: 10.1038/s41598-021-99998-z, www.nature.com/articles/s41598-021-99998-z.

[13] D. H. Kim, W. J. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 223–236, 2019. doi: 10.1109/TAFFC.2017.2695999.

[14] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested LSTM for facial expression recognition," *Neurocomput.*, vol. 317, no. 2, pp. 50–57, 2018. doi: 10.1016/j.neucom.2018.07.028.

[15] J. L. López-Hernández, I. González-Carrasco, J. L. López-Cuadrado, and B. Ruiz-Mezcua, "Framework for the classification of emotions in people with visual disabilities through brain signals," *Neuroinform.*, vol. 15, pp. 642766, 2021. doi: 10.3389/fninf.2021.642766.

[16] K. Okokpujie, S. John, C. Ndujiuba, and E. Noma-Osaghae, "Development of an adaptive trait-aging invariant face recognition system using convolutional neural networks," in *Inf. Sci. App.:*, 2019, pp. 411–420.

[17] X. Li, Z. Yang, and H. Wu, "Face detection based on receptive field enhanced multi-task cascaded convolutional neural networks," *IEEE Access*, vol. 8, pp. 174922–174930, 2020. doi: 10.1109/ACCESS.2020.3023782.

[18] X. Fan, M. Jiang, A. R. Shahid, and H. Yan, "Hierarchical scale convolutional neural network for facial expression recognition," *Cogn. Neurodyn.*, vol. 16, no. 4, pp. 847–858, 2022. doi: 10.1007/s11571-021-09761-3.

[19] S. Singh and F. Nasoz, "Facial expression recognition with convolutional neural networks," in *10th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Las Vegas, NV, USA, 2020, pp. 0324–0328.

[20] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Comput & Applic*, vol. 35, pp. 23311–23328, 2023. doi: 10.1007/s00521-021-06012-8.

[21] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sens.*, vol. 18, no. 2, pp. 401, 2018. doi: 10.3390/s18020401.

[22] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000. doi: 10.1109/34.895976.

[23] F. Sultana, A. Sufian, and P. Dutta, "Evolution of image segmentation using deep convolutional neural network: A survey," *Knowl.-Based Syst.*, vol. 201, pp. 106062, 2020.

[24] C. Chen and F. Qi, "Single image super-resolution using deep CNN with dense skip connections and Inception-ResNet," in *9th Int. Conf. Inf. Tech. Med. Educ. (ITME)*, Hangzhou, China, 2018, pp. 999–1003. doi: 10.1109/ITME.2018.00222.

[25] D. Pathak and U. S. N. Raju, "Content-based image retrieval using GroupNormalized-Inception-Darknet-53," *Int. J. Multimed. Inf. Retr.*, vol. 10, no. 3, pp. 155–170, 2021. doi: 10.1007/s13735-021-00215-4.

[26] X. F. Xu, L. Zhang, C. D. Duan, and Y. Lu, "Research on inception module incorporated siamese convolutional neural networks to realize face recognition," *IEEE Access*, vol. 8, pp. 12168–12178, 2020. doi: 10.1109/ACCESS.2019.2963211.

[27] W. Moungsouy, T. Tawanbunjerd, N. Liamsomboon, and W. Kusakunniran, "Face recognition under mask-wearing based on residual inception networks," *Appl. Comput. Inform.*, vol. 323, no. 8, pp. 707, 2022. doi: 10.1108/ACI-09-2021-0256.

[28] W. Wu, Y. Yin, X. Wang, and D. Xu, "Face detection with different scales based on Faster R-CNN," *IEEE Trans. Cybern.*, vol. 49, no. 11, pp. 4017–4028, 2018. doi: 10.1109/TCYB.2018.2859482.

[29] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *2017 12th IEEE Int. Conf. Automatic Face Gesture Recognit.*, Washington DC, USA, 2017, pp. 650–657.

[30] I. A. Siradjuddin and A. Muntasa, "Faster region-based convolutional neural network for mask face detection," in *2021 5th Int. Conf. Inf. Comput. Sci. (ICICoS)*, Semarang, Indonesia, 2021, pp. 282–286.

[31] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "CMS-RCNN: Contextual multi-scale region-based CNN for unconstrained face detection," in B. Bhanu, A. Kumar (Eds.), *Deep Learning for Biometrics. Advances in Computer Vision and Pattern Recognition*, 2017. doi: 10.1007/978-3-319-61657-5_3.

[32] X. Sun, P. Wu, and S. C. Hoi, "Face detection using deep learning: An improved faster RCNN approach," *Neurocomput.*, vol. 299, no. 2, pp. 42–50, 2018. doi: 10.1016/j.neucom.2018.03.030.

[33] H. Wang, Z. Li, X. Ji, and Y. Wang, "Face R-CNN," arXiv preprint arXiv:1706.01061, 2017.

[34] D. Yi, J. Su, and W. H. Chen, "Probabilistic Faster R-CNN with stochastic region proposing: Towards object detection and recognition in remote sensing imagery," *Neurocomputing*, vol. 459, no. 1, pp. 290–301, 2021. doi: 10.1016/j.neucom.2021.06.072.

[35] C. Wang and Z. Peng, "Design and implementation of an object detection system using Faster R-CNN," in *Int. Conf. Robots Intell. Syst. (ICRIS)*, Haikou, China, 2019, pp. 204–206.

[36] F. Kong, "Facial expression recognition method based on deep convolutional neural network combined with improved LBP features," *Pers. Ubiquitous Comput.*, vol. 23, pp. 531–539, 2019. doi: 10.1007/s00779-019-01238-9.

[37] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, and V. Palade, "Stacked deep convolutional auto-encoders for emotion recognition from facial expressions," in *2017 Int. Joint Conf. Neur. Netw. (IJCN)*, Alaska, USA, 2017, pp. 1586–1593.

[38] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with an attention mechanism"," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, 2018. doi: 10.1109/TIP.2018.2886767.

[39] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, and M. Zareapoor, "Hybrid deep neural networks for face emotion recognition," *Pattern Recognit. Lett.*, vol. 115, no. 2, pp. 101–106, 2018. doi: 10.1016/j.patrec.2018.04.010.

[40] J. Cai, O. Chang, X. L. Tang, C. Xue, and C. Wei, "Facial expression recognition method based on sparse batch normalization CNN," in *2018 37th Chin. Control Conf. (CCC)*, Wuhan, China, 2018, pp. 9608–9613. doi: 10.23919/ChiCC.2018.8483567.

[41] M. Z. Uddin, M. M. Hassan, A. Almogren, M. Zuair, G. Fortino and J. Torresen, "A facial expression recognition system using robust face features from depth videos and deep learning," *Comput. Electr. Eng.*, vol. 63, no. 3, pp. 114–125, 2017. doi: 10.1016/j.compeleceng.2017.04.019.

[42] D. K. Jain, Z. Zhang, and K. Huang, "Multi angle optimal pattern-based deep learning for automatic facial expression recognition," *Pattern Recognit. Lett.*, vol. 139, no. 4, pp. 157–165, 2020. doi: 10.1016/j.patrec.2017.06.025.