EAI Endorsed Transactions
on Internet of Things

Research Article EAI.EU

Innovative Human Interaction System to Predict College Student Emotions Using the Extended Mask-R-CNN Algorithm

Dinesh P*[1], Thailambal G[2]

Department of Computer Science, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India

Department of Advanced Computing and Analytics, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India

Abstract

INTRODUCTION: Recently, there's a growing need for self-decisive and intelligent machines that can understand and respond to human emotions and gestures, particularly among college students, to automate tasks and improve interactions. Facial expression-based emotion recognition is becoming increasingly important in Artificial Intelligence (AI) and computer vision. Traditional manual emotion detection methods are slow, inefficient, and often limited to a few basic emotions. Many studies has focused on object detection systems for effective emotion prediction, but these often suffer from speed, precision, and computational complexity limitations. METHODS: Hence, to address the computational issue and improve object detection performance, the proposed model employs Deep Learning (DL) based adaptive feature spatial anchor refinement with a mask region-based convolutional neural network (Mask RCNN). It is employed to enhance feature extraction with ResNet101, extracts deep features from the input images and detects objects at various scales and aspect ratios. Mask RCNN is a segmentation algorithm, generates a segmentation mask for each detected object. It uses the Facial Expression Recognition (FER) 2013 dataset for the evaluation process. Correspondingly, the efficacy of the projected model is calculated via various evaluation metrics, such as the recall, precision and mean average precision (mAP), to estimate the performance of the proposed DL method. RESULTS: It achieves 0.75298 for MAP@50, 0.70252 for precision and 0.66606 for recall. Furthermore, a comparison of existing models reveals the efficiency of the proposed DL method. CONCLUSION: The present research is intended to contribute to emerging object detection methods for enhancing real-time analysis of student emotions in various environments, such as classrooms and online education.

Keywords: Object detection, emotions prediction, Deep Learning, Mask RCNN, Facial expression recognition

Received on 19 November 2024, accepted on 25 February 2025, published on 10 March 2025

doi: 10.4108/eetiot.7874

1. Introduction

Facial expression recognition (FER) [1] is the term for an inference drawn from the movement of facial features or facial deformations that depend on the facial muscles being activated [2]. Humans possess this special skill by nature. In terms of affective computing, FER has been incredibly popular in recent years [3]. The face is the most expressive aspect of humans and is more efficient at communicating than words and body gestures are.

[*] Corresponding author. Email: dineshlogo02@gmail.com

FER is an essential component in the educational field [4]. Online education has become increasingly popular over the past few decades. Universities and training facilities also provide FER applications. Above all, compared with the old educational system, online education is less constrained and facilitates more effective communication. This naturally makes teachers suspicious of traditional teaching approaches. With the exception of abilities that require accuracy and awareness, students' learning outcomes from e-learning are generally superior [5]. Unquestionably, the rapid development of e-learning will lead to greater flexibility and higher attendance rates, as well as improved convenience. This makes room for expansion in the next period. However, maintaining the highest level of student engagement through improved interaction and focus is crucial to improving the teaching-learning process. Additionally, this enhances the online learning environment [6]. Therefore, the FER2013 datasets with distinct emotions, such as sadness, happiness, fear, disgust, anger, neutrality and disgust, are considered. However, by employing an FER13 and various datasets, the existing methods lack accuracy, which requires more attention [7, 8]. The advantages of deep learning (DL) and machine learning (ML) provide several applications in detection mechanisms. Therefore, the present model addresses the possibility of implementing AI to facilitate the process of object detection [9].

Accordingly, to detect various emotions of the students, the study intended to execute a convolutional neural network (CNN) to predict the facial expressions of the students by employing a CNN that performs better in terms of classification. The FER2013 database is considered in the research process. Research has been executed in three phases. The existing system has achieved an accuracy rate of 70% in the FER2013 dataset [10]. It assists the teacher in identifying the emotions of the students and can construct an enhanced e-learning classroom [11]. The CNN-1 is likely deployed to assess the affective states of 1 student with the availability of a frame with a single image. CNN-2 is intended to identify the numerous students in one image frame. In addition, the prevailing model possesses the ability to reveal the affective state of students in an entire classroom. In addition, the existing model has been employed to increase the teaching and learning processes for both teachers and students. It outperforms the other methods on both posed and spontaneous datasets [12]. Correspondingly, a deep neural network (DNN) has been adopted by the prevailing model for the detection of landmarks that appear in the human face. The hybrid method of the DNN is created for feature extraction, which belongs to the optical flow of micro expression. After the removal of redundant features, followed by improvisation in the features of optical flow, there is an improvement in the accuracy rate [13]. In the same way, the conventional work has deployed an employed supervised learning method that consists of 8 methods, and it is applied to 3 datasets: AffectNet, FER13 [10] and RAF-DB. It has been shown that semisupervised learning makes use of labelled data in a minor portion and that better performance is achieved by employing fully labelled datasets. The hyperparameters that are utilised in semisupervised methods acquire a superior understanding of the optimal settings of the FER dataset [14]. These conventional object detection studies often re on fixed anchor boxes, which can lead to inaccuracies, especially in complex scenarios like facial expression recognition

To overcome inaccuracy and computational complexity limitations, the proposed model uses a certain set of procedures to improve the performance of object detection through the classification of facial features and various emotions. The proposed model uses a publicly available dataset, namely, FER2013. Once the input data are loaded, it processes the pre-processing technique to perform image resizing and other techniques. Then, the pre-processed data are divided into two parts at a ratio of 70:15:15, where 70% is used for the training process, 15% is used for testing, and 15% is used for the validation process. The training set is passed to the proposed DL model of adaptive feature spatial anchor refinement with a mask region-based convolutional neural network (Mask RCNN) to perform object detection. This allows for more precise localization and understanding of the object's shape, which is particularly useful in emotion recognition tasks where facial expressions are subtle and nuanced. Then, the testing and validating data are tested and validated by the proposed modified detection model and evaluated through performance metrics. The major contributions of the present DL model are provided below.

- To utilise a DL-based model for object detection to predict student emotions based on facial reaction with FER2013 dataset images and their annotations.
- To employ adaptive feature spatial anchor refinement with the Mask RCNN to perform object detection.

1.1 Paper Organisation

The flow of the present model is given here: section 1 provides an overview of the background of the proposed model. Section 2 reviews the conventional literature related to object detection and problem identification. Section 3 precisely describes the proposed methodology. Furthermore, section 4 provides a table and graphical representation of the data analysis. Section 5 discusses the results of the proposed DL model compared with those of traditional methods. Finally, section 6 concludes the present model and discusses future research.

2. Literature Review

The present section provides an analysis of various existing studies on object detection along with other techniques for the prediction of emotion classification systems.

Accordingly, in study [15] explored with Convolutional Neural Networks (CNN) and the Viola-Jones algorithm to detect key facial regions (eyes and lips) and recognize seven basic emotion. The prevailing research has developed a multilabel emotion detection architecture (MEDA) for detecting emotions in text pieces. It has been evaluated on an emotional corpus and utilised the NLPCC2018 and RenCECps datasets. It has attained satisfactory performance [16]. In contrast, annotation information from many images is required to train CNN-based object detection methods [17]. The manual annotation cost has been high, and the aircraft targets have usually been small in detecting aircraft from RSIs. Research has demonstrated that AexNet-WSL (weakly supervised learning in AlexNet) can overcome these issues. A better accuracy level was attained by the AlexNet-WSL method [18]. Likewise, the conventional model has developed a stimuli-aware technique that consists of 3 stages, namely, stimulus selection (S), emotion prediction (R) and feature extraction (O), and has been developed for emotion prediction by analysing emotions from various emotional stimuli. It uses 4 visual emotion datasets that achieve 72.42% accuracy [19]. In contrast, the prevailing model has presented a framework that permits social robots to predict various emotions and to save the data in the semantic repository. It is based on an EMONTOlogy (EMONTO). It utilises the SemEVAL-2018 dataset, which achieves an accuracy of 0.535 [20]. It has achieved 93% accuracy on the validation dataset On the basis of the processing of video faces, the conventional model has a pipeline. Initially, clustering methods, tracking and facial detection are used to filter the facial sequences of every student. Experimental results have shown that the conventional model has attained 2% accuracy over the prevailing network [21]. Furthermore, an NN has been utilised to extract the emotional features in all frames and emotion recognition in the Wild (EmotiW) dataset for evaluation [22]. This specific [23, 24] study focused on monitoring the distinct facial expressions of students to create an engaging classroom. The CNN method was proposed in this study to detect the facial expressions of students, such as yawning, sleepiness, frustration, focus and confusion. To determine the efficiency of the model, three publicly available datasets, BAUM-1, YawDD and DAiSEE, are taken into account. The traditional model has obtained accuracy rates of 76.90% and 78.70%, respectively [25]. In contrast, the conventional model offers brisk measures to organise whether the CNN model performs better when it utilises image raw pixel data for training. In addition, whether it is good to offer CNNs some added data, such as HOG and facial landmark characteristics, is unknown. It utilises the FER-2013 dataset and achieves accuracies of 59.1% and 75.1% when trained without and with feature extraction, respectively [26].

Concomitantly, the conventional model employs CNN-LSTM for emotion detection. It has trained and tested on the CREMA-D dataset and the RAVDEES dataset [27], respectively. It has 6 basic emotions: Sad, Fear, Happy, Angry, Neural and Disgust. It has achieved 78.52% and 63.35%, respectively. of accuracy on the CREMA-D dataset and the RAVDEES dataset, respectively [28]. Similarly, the prevailing research has developed the Mask RCNN model, namely, the SCMASK R-CNN. It has been utilised in the detection of remote sensing images with high resolution, which consists of complicated backgrounds and dense targets. The WFA-1400 based on the DOTA dataset [29] was used for an evaluation processThe prevailing approach has presented a real-time method for emotion detection and employment in the applications of robotic vision. Three types of datasets have been utilised, such as the Database of Real-World Affective Faces (RAF-DB), Japanese Female Facial Expression (JAFFE) [30] and Cohn-Kanade (CK+) datasets. The results have shown that the prevailing model has achieved 97% accuracy.

In parallel, the existing work [31] has used 3 commonly utilised facial expression databases and developed manipulation series for stimulating the quality of image deductions. Two sets of experiments were conducted to assess the emotion recognition abilities of 5 commercial recognition systems, namely, Baidu, Face++, Affectiva, Amazon and Microsoft Azure. It has been suggested that the use of manipulation techniques for the subsequent testing of emotion recognition systems could be better. One layer is a fully connected layer along with a rectified linear activation function (ReLU) and a Softmax layer. Experimental results have indicated that the conventional model has attained 92.66% accuracy in mixed datasets;

however, it has attained 94.94% accuracy in cross datasets [32]. Similarly, traditional studies have deployed an architectural design of CNNs for FER mechanisms that consists of 5 convolutional layers [33, 34]. Similarly, the prevailing model was developed to recognise patterns of emotion expression and utilises a DL CNN model [35] with TensorFlow, retraining concepts and Keras. The dataset was collected from a CSV file that was converted to images and used to classify the emotions in expressions. It has achieved a moderate accuracy in emotion prediction task. The study [36] has utilized inception V3 feature extractor with FRCNN segmentation method for face emotion recognition and the findings implicated that the model has yielded the better accuracy. Another study [37] has employed Feature Pyramid Network (FPN) integrated with a Modified Mask R-CNN, using Inception-ResNetV2 to extract multi-level features from video frames. The validation has been performed on the CoCo dataset, and revealed the better accuracy. Similarly, in [38] Mask RFCT segmentation method has been used to generate unique feature vectors for each object instance from video. The algorithm efficacy has been verified on the YouTube-VIS dataset and attained the better accuracy.

2.1 Problem identification

- Conventional research needs to pay more attention to weak and low-source emotion detection and recommended to develop an innovative feature extraction and analysis techniques to enhance performance in terms of accuracy [16].
- Although the prevailing model has achieved better results, it is limited in facial color, as it does not use dark-tone faces and images to detect emotions [32].
- Conventional studies limited with generalizability issue, because skill learned on static image may not translate always perfect outcome in emotion recognition [25].

3. Proposed Methodology

The proposed method identifies and extracts information from a given dataset for object detection. The detection is carried out by implementing adaptive feature spatial anchor refinement with Mask RCNN, a deep learning model. Existing works on object detection have produced inaccurate results with slow convergence speed. Therefore, the proposed model utilises the DL algorithm for the detection of FER data. Moreover, the flow of the proposed DL method is shown in the figure below. 1.

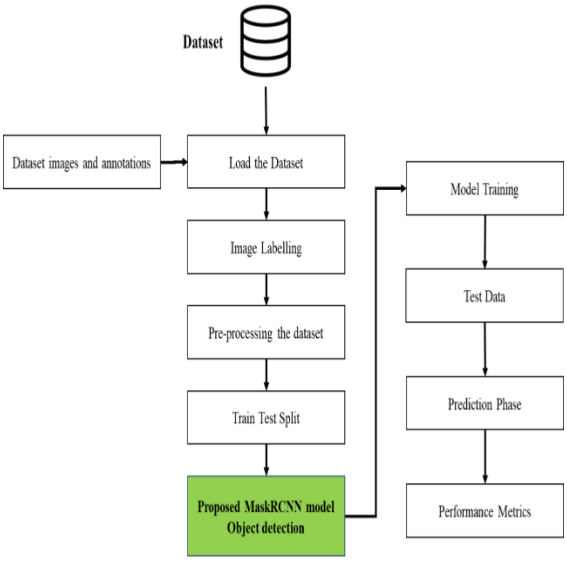

Figure 1. Overall Flow of the Proposed Methodology

Figure 1 shows the proposed model's flow. This signifies the proposed system creation for object detection in a student-based environment. It comprises dataset loading, image labelling, data preprocessing, data splitting and object detection processes via the proposed modified Mask RCNN model. The detection model with the Mask RCNN initially functions with the FER dataset, which is processed in the Python environment. It functions on the Spyder platform, which includes the GUI function. Correspondingly, it functions with a classification model where the efficacy of the classification is calculated via performance metrics.

3.1 Dataset Description

The present method utilises the FER 2013 dataset [39], which is derived from Roboflow. The FER 2013 dataset includes both grayscale facial images, which are 48 × 48 pixels. The faces are automatically registered such that the face is less or more focused and has a similar amount of space in all the images. The task is to classify each facial image on the basis of the emotion that shows facial expression into 7 (0--6) classes and their annotations. These include surprise, disgust, and fear, happy, angry, sad and neutral. Furthermore, the training set and public test set contain 28,709 and 3,589 samples, respectively. The official link of the utilised dataset is provided below.

3.2 Data Preprocessing

Pre-processing is a technique of changing the default data into a proper dataset, which is processed to check the label encoding, feature scaling and other inconsistences before being applied to the algorithm.

In addition, the pre-processing technique improves the detection performance of the proposed method. The pre-processing technique involves a series of consecutive methods that are necessary for generating and storing intermediate files. Managing multiple folders is a challenging task. To address this issue, the directory tree folder model is employed to track the sequential activities of the system. The directory tree for the model includes two dedicated folders for the modified Mask RCNN, as indicated in Table 1. The dataset is organised in the particular format depicted in Table 1.

Table 1 Directory tree structure of the Mask RCNN system

Model	Explanation
Mask RCNN Dataset	• The first subdirectory (train) contains training images located within the subfolder named "images and its xml files." • The second subdirectory (val) contains validation images stored in the subfolder named "images and its xml files." • The third subdirectory (test) contains testing images found in the subfolder named "images and its xml files." • The final three directories (train.xml, val.xml, test.xml) are used for storing specific data files.

The modified Mask RCNN is deployed on the Spyder platform, which is a freely available open-source platform that provides benefits such as simple usage, syntax highlighting, and efficient debugging. It includes a graphical user interface (GUI) that allows users to engage with data and code through familiar interfaces. In this study, the Mask RCNN model implementation employs CPU processing to facilitate training and inference from the system. The first step involves dataset preparation in COCO format with annotations in ‘xml’ format, which is imported to the working directory. Thereafter, a Colab notebook was created in which the model was cloned from the original Mask RCNN repository. Then, all the required dependencies are implemented by executing the requirement.txt file that covers all the libraries needed for processing methods. This file comprises libraries such as NumPy, Keras, OpenCV, python, Scikit image, Pillow, matplotlib, SciPy TensorFlow, h5py, imgaug and IPython.

The next step involved downloading the retrained weights of the modified Mask RCNN and saving it in the original folder of the Mask RCNN repository. For code implementation, some changes are made in the cloned repository to start training on the custom dataset. The steps are described below.

(i) Create a folder named Dataset inside the main repository.

$$D{:}\backslash Users\backslash Desktop\backslash Dinesh\_code \\ \backslash Dataset.$$

(ii) Inside ‘Dataset’ -----> create three subfolders as $‘train’, ‘test’\ and\ ‘val’.$

(iii) All the images labelled for validation, training and testing are copied into their respective folders along with the ‘xml’ file.

3.3 Data splitting

In deep learning, data splitting is used to eradicate data overfitting. Essentially, DL uses the data splitting technique to train the respective model where the training data are added to the proposed method for equipping the training stage parameters. After the training process, the test set data are measured to calculate the present model for handling the observations. In the present model, the original data are split into 3 sets at a ratio of 70:15:15, which indicates that 70% of the data are utilised for training, 15% of the data are applied for validation, and the remaining 15% of the data are utilised for testing to calculate the performance of the respective technique.

3.4 Object Detection - Adaptive Feature Spatial Anchor Refinement with Mask RCNN

The traditional Mask RCNN was developed on the basis of Faster R-CNN, which has two outputs for each candidate object: a class label and a bounding box offset. In contrast, the Mask RCNN is an advanced version of Faster R-CNN specifically for object segmentation in images and videos. By including a branch dedicated to predicting an object mask (region of interest) alongside the existing branch for bounding box detection. It introduces a third branch that generates the object mask, this additional mask output is separate from the class and box outputs, allowing for a more detailed spatial layout of an object to be extracted. It produces a segmentation mask for every identified object, enhancing accuracy in locating and understanding the shape of the object. This is especially helpful in tasks like emotion recognition, where subtle facial expressions need to be detected.

The proposed model uses adaptive feature spatial anchor refinement with the Mask RCNN for effective object detection, especially for recognising the facial emotions of students. It leverages the ability of the modified Mask RCNN for instance object detection of facial expressions. The Mask RCNN is known as an extension of the Faster R-CNN mechanism through the addition of Masks for detecting segmentations in regions of interest (RoIs). It not only detects objects and movements but also delineates detailed boundaries. Likewise, spatial anchors are predefined bounding boxes that facilitate object detection at various scales and aspect ratios within DL frameworks. Traditional methods often use fixed anchor boxes, which may not be suitable for objects with varying sizes and shapes. To address this limitation, an adaptive mechanism can be used to refine anchors based on features extracted from input images. This process involves feature extraction using a backbone network like ResNet101, followed by dynamic adjustment of anchor boxes to focus on relevant areas. Additionally, an attention mechanism can be employed to weigh the importance of different features, improving the model's ability to detect subtle variations. It is certainly used in facial emotion recognition, thereby effectively recognising facial features and determining emotional states. The process of facial detection starts with detecting facial images in the data source via the Mask RCNN. It involves searching for key facial movements and classifying the face from its background. Once facial features are detected, a DL model such as a CNN is employed to precisely examine the features. Adaptive feature spatial anchor refinement with the modified Mask RCNN improves the ability of the proposed model to centre on relevant features by refining the spatial anchors, which are often based on adaptive learning. Dynamically, it allows the present model to adjust attention, thereby enhancing accuracy in identifying subtle facial expressions. The following figure is shown. 2 describes the mechanism of the Mask RCNN.

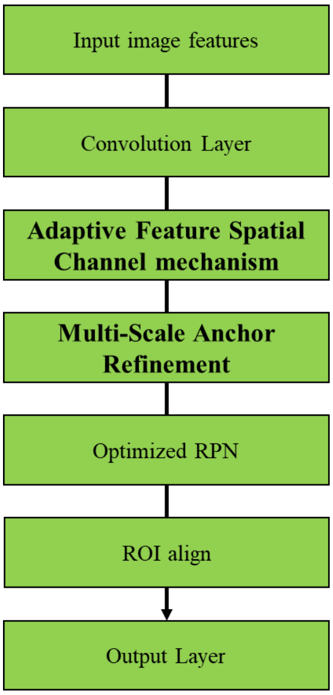
Input image features
Convolution Layer
Adaptive Feature Spatial Channel mechanism
Multi-Scale Anchor Refinement
Optimized RPN
ROI align
Output Layer

Figure 2 Mechanism of the Mask RCNN

The dilated convolution set along a certain size is added by the object detection model modified Mask RCNN to enhance the instance detection impact, and a ResNet101 backbone network is used to gather more discriminative feature information. Using a neural network to create a region proposal immediately is the fundamental concept of the RPN. The RPN receives the convolutional feature map as input from the convolutional neural network. The Mask RCNN was built via Faster R-CNN. While Faster R-CNN has 2 outputs for each candidate object, a class label and a bounding-box offset, Mask RCNN is the addition of a third branch that outputs the object mask. The added mask output is distinct from the box and class outputs and requires the extraction of a much finer spatial layout of an object. Thus, some of the advantages of the Mask RCNN are the flexibility and efficacy of the model. In a traditional CNN model, the 3×3-sized convolution kernel is utilised to combine the feature data. $3\ x\ 3$ Kernel size results in limited spatial data because of the convolution kernel size limitation, thereby reducing the amount of information in a large view. Data loss cannot have a crucial effect on natural facial image recognition. Initially, an average pooling function with a size of $r \times r$ is used to acquire spatial information T_1 via equation (1):

$$T1 = AvgPoolingr(M_1) \quad (1)$$

Second, T_1 is passed by K_2 convolution and upsampling in the sequences, after which it performs elementwise summation in M_1 before being sent to the sigmoid function. Therefore, the output achieves elementwise multiplication along with T_2 acquired through K_3 convolution with M_1, as shown in equation (2):

$$N_1' = (M_1 * K_3) \bullet \sigma(M_1 + Up(T_1 * K_2)) \qquad (2)$$

where σ and $*$ denote the sigmoid function and convolution, respectively. Third, N_1 is acquired through convolution K_4, as shown in equation (3):

$$N_1 = N_1' * K_4 \qquad (3)$$

To improve the relationship between contexts, the modified Mask RCNN notably increases the intensity of self-calibration convolution in high-level feature maps. This makes each spatial position more informative and improves the ability to extract low-level feature information from the feature maps (such as a clearer texture). The key of the RPN is to utilise a neural network to create the region directly. The convolutional feature map returned through the CNN is utilised as the RPN input. The RPN is utilised to optimise and reduce the number of calculations, enhancing accuracy and training speed. Two convolutional kernels of size 1×1 are utilised for parallel convolutional functions. The detection layer is utilised to detect the target area and its background. Bounding box regression performs a primary correction on the target boundary box, as shown in equation (4):

$$d(P) = \omega^T M(P) \qquad (4)$$

where M_p is the feature map feature vector, $d(P)$ denotes the detected region, and ω^T denotes the parameter matrix. Figure 3 shows the Mask RCNN model.

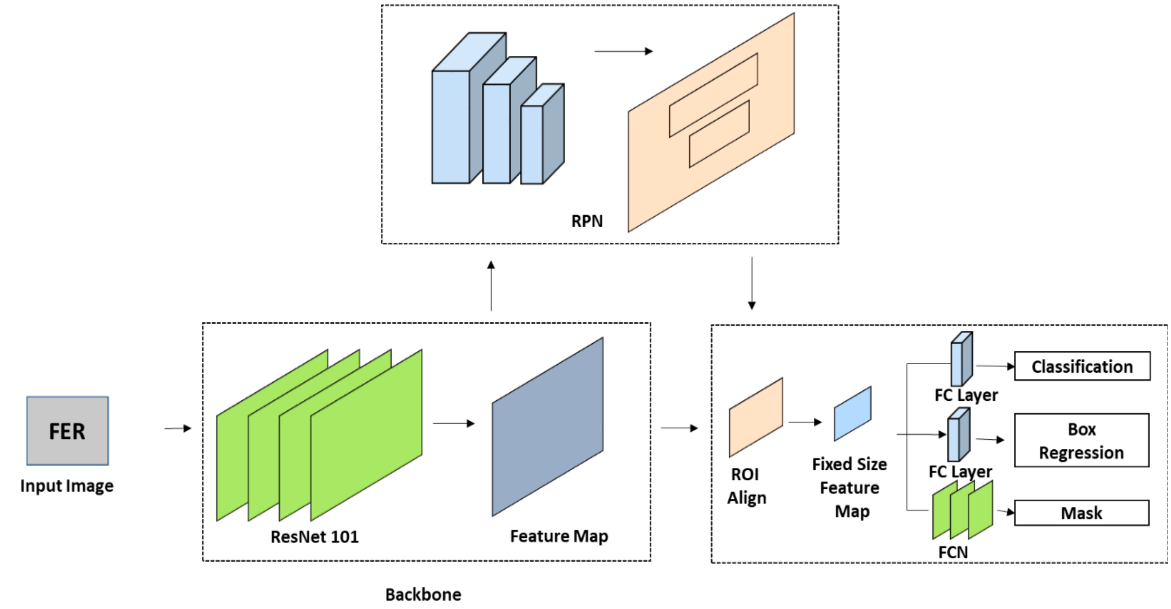

Figure 3. Modified Mask RCNN

Figure 3 shows that the architecture of the modified Mask RCNN is separated into four major parts, as described below.

- Backbone
- RPN (region proposal network)
- Bounding box segmentation and regression
- ROI (Region of interest)

The modified Mask RCNN method is designed to reduce the complexity of convolution by incorporating three key features. First, it employs two additional classifiers within the network to address the issue of gradient vanishing. Second, it utilises a wider and deeper network architecture and updates the ResNet101 module with 3 × 3 convolutions to address the problem of information loss. The RPN is a binary classification network that categorises images into two groups: background and targeted objects. It identifies one or more regions containing the desired objects, which are then used as input in the ROI alignment process. These regions are subsequently processed to obtain a fixed-size feature map. The ROI alignment method classifies and locates each ROI by employing a bilinear interpolation model in the feature map process, ensuring spatial correspondence with every pixel in the data. Correspondingly, every ROI in the system is altered to a fixed size feature map where the input data are divided into two branches, where one branch functions with targeted object recognition and bounding box classification. Similarly, a fully connected layer is utilised to classify a particular category. In the same way, a full convolution layer is used, which produces a mask with a similar shape and size.

4. Results and Discussion

The current section presents the results obtained by the proposed model on multiclass object segmentations. It illustrates the EDA, experimental results, comparative results and discussion of the proposed model outcomes.

4.1 Exploratory Data Analysis (EDA)

The EDA is used to overview and analyse the images in the dataset. Figure 4 shows the image data from the FER data source.

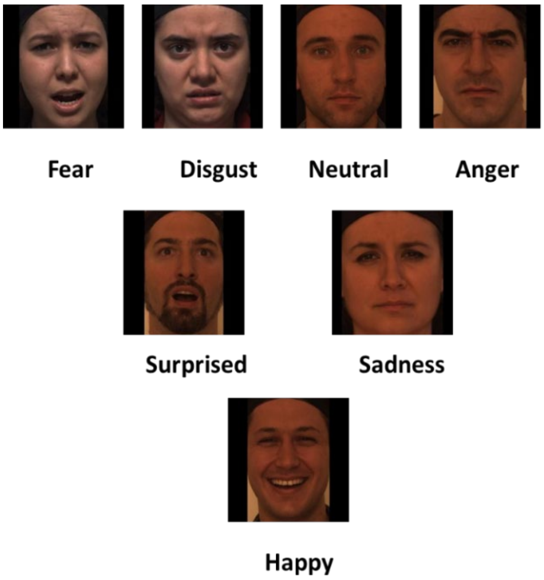

Figure 4. Images from the FER Data Source

Figure 4 represents the sample data contained in the FER 2013 dataset. The dataset comprises 0 to 6 classes, such as Fear, Disgust, Neutral, Anger, Surprised, Sadness and Happy. It can be labelled as 0-angry, 3- happy,2-fear,5 –Disgust,6-surprise, 1-neutral, 4-sad

4.2 Performance Metrics

The section signifies the metrics used in the respective model to analyse the effectiveness in the classification of objects. Accordingly, the performance metrics used in the presented system are precision, recall, and mean average precision (MAP).

4.2.1 Precision

The precision is the performance metric that signifies the amount of positive prediction accomplished in the projected system. It is used to examine the efficacy of the proposed classification model. The formula for precision [40] is depicted in equation (5):

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

(5)

Where, TP and FP are true positive and false positive, respectively.

4.2.2 Recall

The recall metric is used to examine the percentage of data that are correctly predicted in the classification. The formula for the recall is shown in equation (6):

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

(6)

Where, TP and FN are true positive and false negative, respectively.

4.2.3 Accuracy

The accuracy metric is an important metric utilised to analyse the number of predictions that are accurately correct in the classification. The formula for accuracy is described in equation (7),

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$$

(7)

Where, TP, TN, FP, and FN are true positive, true negative, false positive and false negative, respectively.

4.2.4 Mean average precision

The average precision (AP) [41], which is the area under the precision-recall curve, is determined separately for each class because precision and recall cannot fully evaluate a model's performance. mAP is the average precision over all the recall values across all the IoUs for the prediction and ground truth thresholds between 0.5 and 0.95.

$$\text{mAP} = \frac{1}{\text{n}} \sum_{k=1}^{k=n} AP_k$$
(8),

4.3 Hardware and Software Requirement

While implementing an emotion detection model there is a specific requirement in hardware and software component. It can be illustrated in table 2,

Table 1 Hardware and Software Components Requirement

Category	Details
Hardware Technologies (Required Hardware details)	**Processing Units**: NVIDIA RTX 3060 or higher for optimal performance.
	Memory: Minimum 16 GB RAM
	Storage: At least 256 GB of SSD space for datasets and model checkpoints.
Software Technologies (Software Frameworks)	**Frameworks**: TensorFlow for model training and deployment.
	Data Processing Tools: OpenCV for image processing, scikit-learn for additional preprocessing tasks

From the table 2, these hardware and software specifications ensure that the emotion detection system can effectively process facial expressions. A typical system might use a camera or web camera to capture face reaction, the model required NVIDIA RTX 3060 or higher GPU for optimal performance. A minimum of 16 GB of RAM is recommended and 256 GB of SSD space should be require for storage purpose, along with software to analyse the input and detect emotions.

4.4 Performance Analysis

The performance of the proposed DL model is considered via several metrics, such as precision, mAP and recall. This section discusses the results accomplished by the proposed mechanism for predicting Disgust, Angry, Fear, Happy, Surprise, Sad and Neutral with the Kaggle-FER 2013 dataset. Figure 5 shows the ROC curve for the proposed Mask RCNN method.

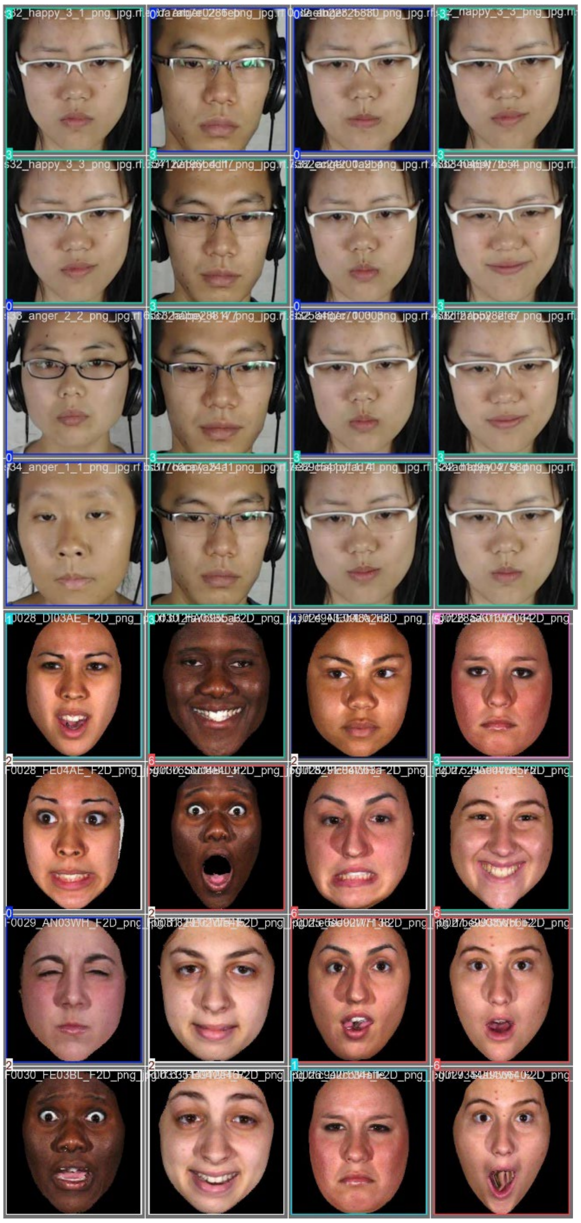

Figure 5 Validation Batch Labels

From figure. 5, the numbers overlaid on the images could indicate the confidence scores assigned by an emotion detection model. The digits might represent the likelihood or intensity of a particular emotion associated with each expression. It was labelled as follows 0-angry, 1-neutral, 2 -fear, 3- happy, 4-sad 5 –Disgust, 6-surprise. The variations in facial expressions across rows and columns might be used to classify different emotional states.

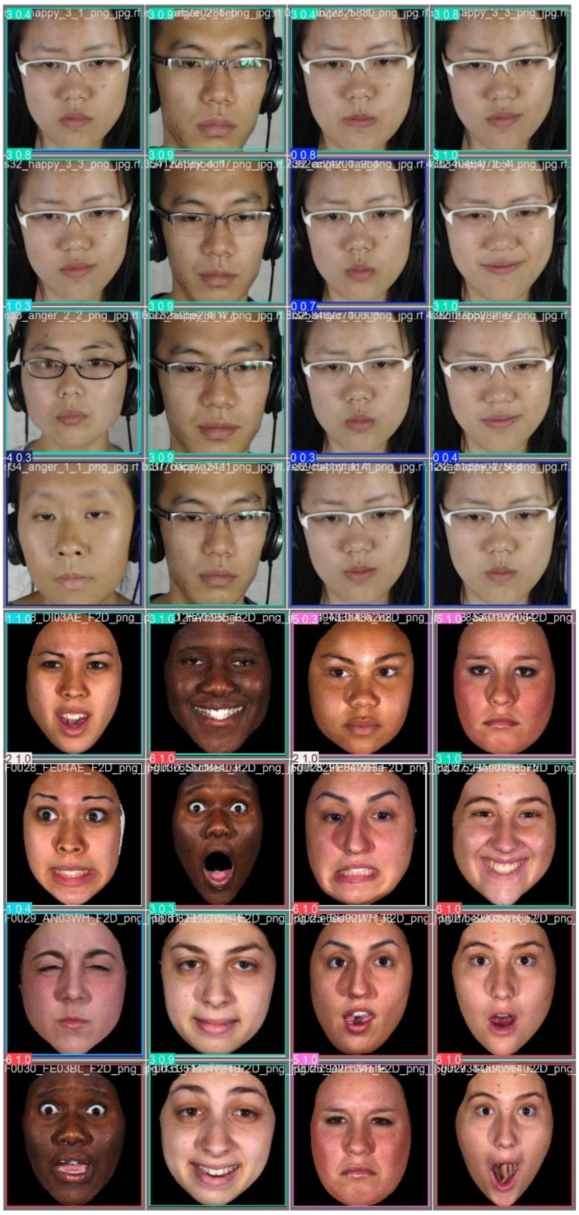

Figure 6 Validation prediction labels

In figure. 6, there are significant differences in the ways in which each person expresses their emotions, and each face in the grid presents a unique emotional expression. Higher numbers may suggest a stronger expression of the related emotion. The numbers may also show how strongly an emotion is felt. One may utilise the color-coded boxes surrounding the faces to show intensity or distinguish between different emotions. These types of models can be trained or tested on a vast range of expressions and facial traits. Furthermore, Figure 7 shows the various metric results for the proposed modified Mask RCNN model.

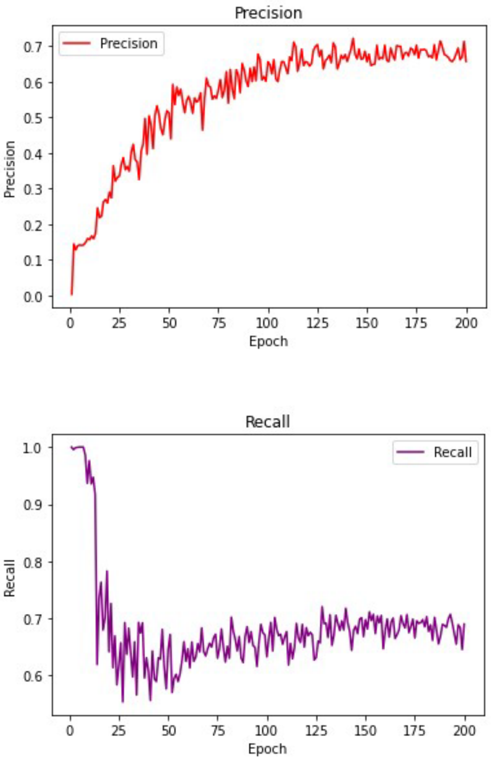

Figure 7 Precision and Recall

From figure. 7 Precision starts low and gradually increases over time. It stabilises at approximately 0.65–0.7 from approximately epoch 100 onwards. Precision measures the proportion of properly predicted true positives to the total number of predicted positive observations, such as true positives and false positives. The proposed DL model improves over time, suggesting that it becomes more selective about the positive class as training progresses. Likely, recall estimates the number of actual positives that are properly identified by the present model. The sharp drop in the early epochs indicates that initially, the proposed model was overfitting (trying to capture everything), leading to high recall, but as training progresses, it balances better and focuses more on precision. The fluctuations suggest instability in the present model's ability to consistently identify all positives.

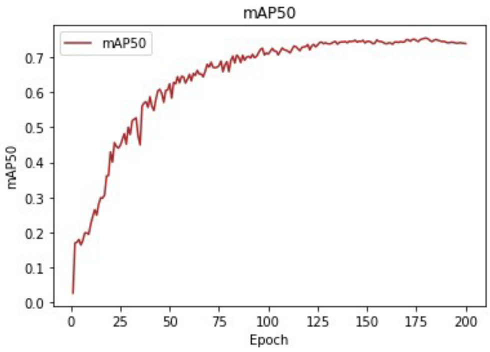

Figure 8 mAP50

Figure 8 indicates that mAP50 reflects the average precision for the task of object detection, where predictions are measured correctly if they overlap at least 50% with the bounding boxes. The above graph shows stable improvement over time, indicating that the present model becomes better at detecting objects with high accuracy as it trains.

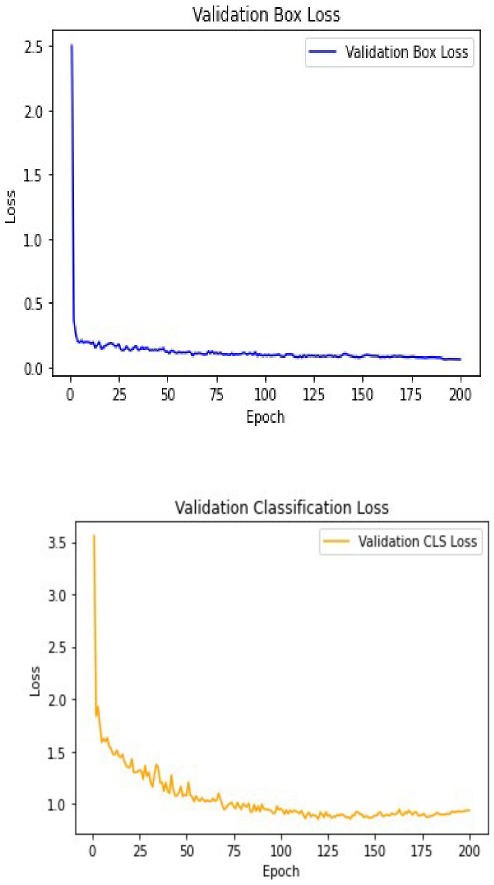

Figure 9 Validation Box and Classification Loss

From figure. 9, the validation classification loss is high (approximately 3.5) and drips suddenly in the first few epochs, which is estimated as the present model quickly learns during early training. After 50 epochs, the loss stabilises at approximately 1.0, showing a diminishing improvement. The final few epochs show a slight upwards trend, indicating potential overfitting where the performance of the present model on the validation set may deteriorate. Moreover, in the validation box loss, a decrease is observed in the first few epochs, suggesting that the present model rapidly learns to regulate the bounding boxes to better match the ground truth. After 20 epochs, the loss stabilises, fluctuates slightly and then shows no significant reduction. This finding indicates that the present model converges and no longer makes crucial improvements in bounding box prediction during validation. When the loss approaches zero, the model achieves high accuracy in bounding box prediction.

4.5 Comparative Analysis

This section presents an external comparison analysis of the proposed DL model in accordance with various performance metrics. The table. 3 shows the comparison of the performance of the existing model with that of the proposed DL model.

Table 2 Comparison of Existing Model [40]

Model	MAP@50	Precision	Recall
YOLOv5s	0.7191	0.5936	0.7189
YOLOv51	0.7097	0.6467	0.6476
YOLOv5x	0.7115	0.648	0.6567
YOLOv7	0.5587	0.4603	0.7251
YOLOv7-tiny	0.66	0.5575	0.7288
YOLOv7x	0.7227	0.6527	0.6944
YOLOv8n	0.7268	0.5674	0.7173
YOLOv8s	0.7285	0.613	0.6961
YOLOv8m	0.7276	0.5656	0.6741
YOLOv81	0.7287	0.6155	0.7187
YOLOv8x	0.7327	0.636	0.7092
YOLOv9c	0.7305	0.6773	0.8112
YOLOv9e	0.7337	0.684	0.8562
Proposed model	**0.75298**	**0.70252**	**0.66606**

The table. 3 shows that the proposed Mask RCNN model achieves better results than conventional methods do. Compared with other conventional models such as YOLOv5s, YOLOv51, YOLOv5x, YOLOv7, YOLOv7-tiny, YOLOv7x, YOLOv8n, YOLOv8m, YOLOv81, YOLOv8x, YOLOv9e and

YOLOv9c, it achieves 0.75298 MAP@50, 0.70252 precision and 0.66606 recall.

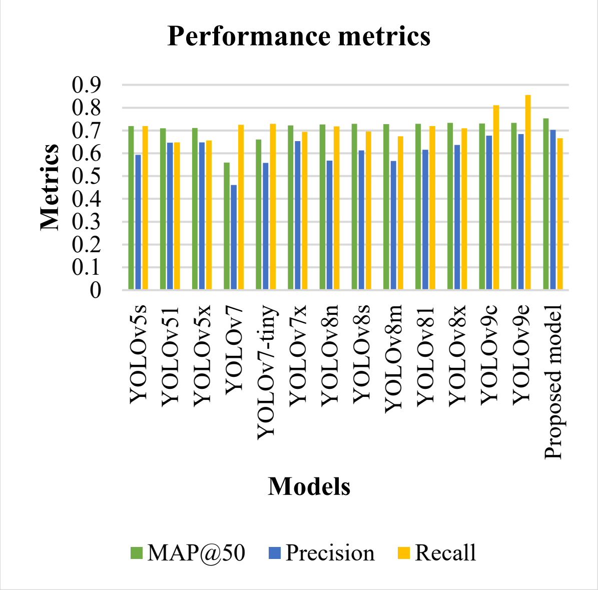

Figure 10 Metrics Comparison [40]

Figure. Figure 10 shows a comparison of the metrics of the proposed model with those of existing models. The proposed model achieves better results, which shows its efficiency in detecting objects in the facial image dataset. The other conventional models yield mAP@50 values of 0.7191, 0.7097, 0.7115, 0.5587, 0.66, 0.7227, 0.7268, 0.7285, 0.7276, 0.7287, 0.7327, 0.7305 and 0.7337 for YOLOv5s, YOLOv51, YOLOv5x, YOLOv7, YOLOv7-tiny, YOLOv7x, YOLOv8n, YOLOv8s, YOLOv8m, YOLOv81, YOLOv8x, YOLOv9c and YOLOv9e, respectively. These values are less than those of the proposed modified Mask RCNN model.

4.6 Discussions

Various datasets have been utilised in conventional models for object segmentation and detection. However, traditional models lack the ability to detect multiple objects from a single model, which hinders the effectiveness of classification. To address this limitation, a proposed method has been developed that incorporates diverse objects such as facial emotions. The accuracy of the classification model plays a crucial role in evaluating its effectiveness. Unfortunately, many conventional models struggle to achieve high accuracy. In contrast, the respective research has surpassed previous studies by achieving an accuracy value that is significantly higher than the average accuracy of existing systems. This remarkable achievement is made possible through the integration of adaptive feature spatial anchor refinement with the Mask RCNN. By incorporating different objects into the FER 2013 dataset, the proposed system has demonstrated superior performance in terms of the object detection mechanism. The results of the proposed research showed a significant enhancement in the accuracy of emotion detection by utilizing adaptive feature spatial anchor refinement with Mask RCNN. This approach contributes to a more robust framework for recognizing subtle emotional cues in facial expressions. Our model sets a new standard for multiple object detection in a single shot by improving precision and recall rates in emotion detection. In academic environments, this model has the potential to be incorporated to track student emotions in real-time. This enables educators to pinpoint disengaged or struggling students and provide timely assistance to improve learning experiences. By examining students' emotional reactions, educators can customize their methods to better cater to individual learners and create a more positive and supportive classroom setting.

5. Practical application of the proposed work

The practical significance of these findings is illustrated by the following real-world case studies:

5.1 Case study1: Emotion detection in class room

In a middle school classroom, our emotion detection model monitors student emotions in real-time, alerting the teacher to signs of confusion or frustration. This allows the teacher to adjust her approach, provide additional explanations, or engage students in discussions, leading to increased student engagement and improved understanding of the material

5.2 Case study 2: Online learning platform

Online learning platforms are using emotion detection to gauge student engagement during lectures. By analyzing facial expressions, instructors receive real-time feedback, enabling them to adjust content for better interaction. This adaptive approach leads to higher completion rates and improved learner satisfaction.

5.3 Case study 3: Mental Health Monitoring

It supports mental health initiatives in schools. By monitoring student emotions during counseling, these models aid counselors in identifying students needing extra support, flagging patterns of emotional distress for follow-up. This leads to enhanced mental health support and early intervention

6. Ethical consideration

In the proposed study, ethical considerations are prioritized throughout the development and implementation of emotion detection methodologies. Accordingly, obtained informed consent from all participants (or their guardians), providing comprehensive details regarding data collection procedures, research objectives, and data usage policies. Data security was paramount, with robust encryption and secure storage protocols employed to prevent unauthorized access. Participants retained ownership and control over their data, with the right to access, modify, or request deletion. Acknowledged the limitations of emotion detection systems as interpretations of emotional signals, rather than direct reflections of internal states. Recognizing the importance of contextual validity, we rigorously tested and validated our system within the specific educational context. Finally, the work committed to transparency by clearly communicating the system's limitations to all stakeholders such as such as educators, mental health professionals, administrators, and technology developers and establishing safeguards against potential misuse or misinterpretation of results.

7. Conclusion

The improvement of an object detection model to predict college students' emotions is highly important in various dimensions, especially in educational settings. Numerous manual methods for detecting emotions are focused on few basic emotions, which are considered to be time-consuming and limited. Therefore, an effective object detection model for predicting student emotions is important for ensuring engagement and learning experience and facilitating decision-making. To resolve this problem, the proposed research employed a DL approach called adaptive feature spatial anchor refinement with the Mask RCNN, which aims to overcome the prediction accuracy for object detection in facial expressions. The FER2013 dataset is utilised in the proposed research to determine the effectiveness of the present model. The present research achieved 0.75298 mAP@50, 0.70252 precision and 0.66606 recall. Constantly, the outcome of the comparative analysis indicated that the respective model has overtook the existing research.

This current research contributes to the vast applications such as middle school class room, online learning platform and metal health monitoring. In classrooms, real-time monitoring of student emotions allows teachers to adjust their approach, leading to increased engagement and understanding. Online learning platforms benefit from emotion detection by providing instructors with feedback on learner engagement, improving completion rates and satisfaction. Mental health initiatives in schools can be enhanced through emotion detection, enabling early intervention and support for students in need. Future studies could explore sophisticated techniques and integrating multimodal data sources like audio and text with visual data to enhance emotion detection model robustness across diverse demographics. In addition, facial images from IoT devices are used to detect emotions for effective results in the object detection model.

Declaration

Conflict of Interest

There is no conflict of interest.

Funding Support

There is no funding support for this study.

Data Availability Statement

Not Applicable.

References

[1] Haq, H., et al., Enhanced real-time facial expression recognition using deep learning. 2024. 3(1): p. 24-35.

[2] Russell, P.W., X. Min, and S.S. Lily, Facial recognition and emotion detection in environmental installation and social media applications, in Encyclopedia of Computer Graphics and Games. 2024, Springer. p. 694-703.

[3] Ekundayo, O. and S. Viriri, Facial Expression Recognition: A Review of Methods, Performances and Limitations.

[4] Li, S. and W. Deng, Deep facial expression recognition: A survey. IEEE transactions on affective computing, 2020. 13(3): p. 1195-1215.

[5] Ouherrou, N., et al., Comparative study on emotions analysis from facial expressions in children with and without learning disabilities in virtual learning environment. Education and Information Technologies, 2019. 24(2): p. 1777-1792.

[6] Wang, W., et al., Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation. Complexity, 2020. 2020: p. 1-9.

[7] Wang, K., et al., Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 2020. 29: p. 4057-4069.

[8] Borgalli, R.A. and S. Surve, Deep Convolution Neural Networks for Cross-Dataset Facial Expression Recognition System.

[9] Yalew, B., Emotion Recognition from Facial Expression Using Convolutional Neural Network. 2024, St. Mary's University.

[10] Gera, D., et al., Dynamic adaptive threshold based learning for noisy annotations robust facial expression recognition. 2024. 83(16): p. 49537-49566.

[11] Lasri, I., A.R. Solh, and M. El Belkacemi, Facial Emotion Recognition of Students using Convolutional Neural Network.

[12] TS, A. and R.M.R. Guddeti, Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. Education and information technologies, 2020. 25(2): p. 1387-1415.

[13] Pei, J. and P. Shan, A Micro-expression Recognition Algorithm for Students in Classroom Learning Based on Convolutional Neural Network. Traitement du Signal, 2019. 36(6).

[14] Roy, S. and A. Etemad, Analysis of Semi-Supervised Methods for Facial Expression Recognition. arXiv preprint arXiv:2208.00544, 2022.

[15] Ali, M.F., M. Khatun, and N.A. Turzo, Facial emotion detection using neural network. the international journal of scientific and engineering research, 2020.

[16] Deng, J. and F. Ren, Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. IEEE Transactions on Affective Computing, 2020. 14(1): p. 475-486.

[17] Yang, K., et al., Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets. The visual computer, 2021. 37: p. 1447-1466.

[18] Wu, Z.-Z., et al., Convolutional neural network based weakly supervised learning for aircraft detection from remote sensing image. IEEE Access, 2020. 8: p. 158097-158106.

[19] Yang, J., et al., Stimuli-aware visual emotion analysis. IEEE Transactions on Image Processing, 2021. 30: p. 7432-7445.

[20] Graterol, W., et al., Emotion detection for social robots based on NLP transformers and an emotion ontology. Sensors, 2021. 21(4): p. 1322.

[21] Wu, Q., et al., Improved mask R-CNN for aircraft detection in remote sensing images. Sensors, 2021. 21(8): p. 2618.

[22] Savchenko, A.V., L.V. Savchenko, and I. Makarov, Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. IEEE Transactions on Affective Computing, 2022. 13(4): p. 2132-2143.

[23] Siam, A.I., et al., Deploying machine learning techniques for human emotion detection. Computational intelligence and neuroscience, 2022. 2022(1): p. 8032673.

[24] Hans, A.S.A. and S. Rao, A CNN-LSTM based deep neural networks for facial emotion detection in videos. International Journal of Advances in Signal and Image Sciences, 2021. 7(1): p. 11-20.

[25] Pabba, C. and P. Kumar, An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. Expert Systems, 2022. 39(1): p. e12839.

[26] Amal, V., S. Suresh, and G. Deepa. Real-time emotion recognition from facial expressions using convolutional neural network with Fer2013 dataset. in Ubiquitous Intelligent Systems: Proceedings of ICUIS 2021. 2022. Springer.

[27] Hazra, S.K., et al., Emotion recognition of human speech using deep learning method and MFCC features. Radioelectronic and Computer Systems, 2022(4): p. 161-172.

[28] Kumar, S., et al., Multilayer Neural Network Based Speech Emotion Recognition for Smart Assistance. Computers, Materials & Continua, 2023. 75(1).

[29] Wu, W., H.-S. Wong, and S. Wu, Pseudo-Siamese Teacher for Semi-Supervised Oriented Object Detection. IEEE Transactions on Geoscience and Remote Sensing, 2024.

[30] Saeed, V.A., A framework for recognition of facial expression using HOG features. International Journal of Mathematics, Statistics, and Computer Science, 2024. 2: p. 1-8.

[31] Manalu, H.V. and A.P. Rifai, Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm. Intelligent Systems with Applications, 2024. 21: p. 200339.

[32] Qazi, A.S., et al., Emotion detection using facial expression involving occlusions and tilt. Applied Sciences, 2022. 12(22): p. 11797.

[33] Meena, G., et al., Identifying emotions from facial expressions using a deep convolutional neural network-based approach. Multimedia Tools and Applications, 2024. 83(6): p. 15711-15732.

[34] Chowdary, M.K., T.N. Nguyen, and D.J. Hemanth, Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing and Applications, 2023. 35(32): p. 23311-23328.

[35] Ganesan, P., et al., Deep Learning-based Interactive Dashboard for Enhancing Online Classroom Experience through Student Emotion Analysis. IEEE Access, 2024.

[36] Angel, J.S., et al., Faster Region Convolutional Neural Network (FRCNN) Based Facial Emotion Recognition. Computers, Materials & Continua, 2024. 79(2).

[37] Yadav, A. and E. Kumar, Object Detection on Real-Time Video with FPN and Modified Mask RCNN Based on Inception-ResNetV2. Wireless Personal Communications, 2024: p. 1-26.

[38] Imran, A., et al., FaceEngine: A Tracking-Based Framework for Real-Time Face Recognition in Video Surveillance System. SN Computer Science, 2024. 5(5): p. 609.

[39] FER2013 expression Computer Vision Project.

[40] Parambil, M.M.A., et al., Navigating the YOLO landscape: A comparative study of Object detection models for Emotion Recognition. 2024.

[41] Mukhiddinov, M., et al., Masked face emotion recognition based on facial landmarks and deep learning approaches for visually impaired people. Sensors, 2023. 23(3): p. 1080.