ISSN: 2093-5374 / E-ISSN: 2093-5382

Student Engagement Analysis using Bi-Up Sampling Feature Pyramid Network with Inter Cross Coordinate Self Attention

A. Naveen[1*], I. Jeena Jacob[2], and Ajay Kumar Mandava[3]

[1*]Scholar, Department of Computer Science and Engineering, GITAM School of Technology, Bengaluru Campus, India. a.naveen21@gmail.com, https://orcid.org/0000-0001-6273-2520

[2]Professor, Department of Computer Science and Engineering, GITAM School of Technology, Bengaluru Campus, India. ijacob@gitam.edu, https://orcid.org/0000-0001-6706-1017

[3]Associate Professor, Department of Electrical, Electronics and Communication Engineering, GITAM School of Technology, Bengaluru Campus, India. amandava@gitam.edu, https://orcid.org/0009-0001-8902-6914

Received: October 24, 2024; Revised: December 10, 2024; Accepted: January 13, 2025; Published: March 31, 2025

Abstract

Facial expressions are physical changes that reflect an individual's feelings, emotions, intentions, or social interactions. In the field of computer vision, facial expression analysis needs a higher level of knowledge. To increase student involvement in class, in recent years increased in employing technology to track and evaluate students' facial expressions. Facial expressions, as a nonverbal form of communication, can provide valuable insights into students' emotional states. This research article aims to explore the potential of class engagement monitoring using facial expressions. Hence, we proposed a Bi-upsampling Feature Pyramid Network (BiusFPN) with multiple attention map integration. We used ResNet18 as a backbone and we also introduced an inter-coordinate attention model which improves the feature extraction from local spatial extended mode with different coordinate representations. This engagement analysis model precisely detects the facial changes and yields an accurate outcome. We integrate both Channel and spatial attention mechanisms to fuse different attention maps for the final representation. The result of the classification layer will be Disengaged or partially engaged or engaged. This approach achieves 68.16% accuracy for the DAiSEE dataset and 83% for the WACV dataset. In this way, the proposed method identifies the facial expression that contributes to the findings of classroom coordination.

Keywords: Machine Learning, Deep Learning, Facial Expression.

1 Introduction

Class monitoring using facial expressions is a technology that uses computer vision, machine learning, and deep learning algorithms to analyze the student's expressions during online or in-person classes. This can be done using cameras or other devices that capture images of students' faces and interpret emotions like engagement, boredom, frustration, or confusion in real time (Abedi & Khan, 2021; Eskandarian et al., 2016; Tamannaeifar & Hesampour, 2016; Amraee & Koochari, 2014). The collected data can be used to provide teachers with insights into how students are engaging with the material, and

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 16, number: 1 (March), pp. 134-153. DOI: *10.58346/JOWUA.2025.I1.008*

*Corresponding author: Scholar, Department of Computer Science and Engineering, GITAM School of Technology, Bengaluru Campus, India.

whether they may need additional support or resources. Additionally, this technology can also be used for attendance tracking, to monitor students' attention and participation in class (Soozanyar & Jafarzadeh, 2017). It's important to note that this technology raises privacy concerns as it's capturing sensitive personal data and its use should be under strict guidelines and regulations (Hu et al., 2018).

The student engagement analysis during class can be implemented by using different features like facial expression recognition, head pose detection, body language analysis, etc (Pravin Kumar et al., 2023). Studies have used the facial expressions approach to analyze engagement in online and in-person classes and have shown that it can be effective in detecting engagement levels (Lemay et al., 2018). Studies have used this approach to analyze engagement in online and in-person classes, and have shown that it can be effective in detecting engagement levels. (Jiang & Jia, 2011) used a dataset of 1,000 students and achieved an accuracy rate of 85.8% in detecting engagement levels. Another approach is to use speech analysis to detect engagement levels. This can include analyzing the pitch, tone, and volume of a student's speech, as well as the content of their speech. The next approach is to use head pose detection or gaze to detect engagement levels. Studies have used this approach to analyze engagement in online and in-person classes, and have shown that it can be effective in detecting engagement levels. This can include analyzing the direction and duration of a student's gaze, as well as the number of times they look away from the screen. In one work (Uçar & Özdemir, 2022), where the authors proposed a system that uses gaze tracking to detect engagement levels in online classes and demonstrated its effectiveness in detecting engagement levels in a virtual classroom setting. Some (Batra et al., 2022; Huang et al., 2018) authors proposed a head pose detection method to assess the engagement level of students during lectures. They used a dataset of 600 students and achieved an accuracy rate of 92.2% in detecting engagement levels. Another approach (Bourel et al., 2001) used gaze tracking to detect student engagement in the classroom (online) by using a deep learning algorithm and found that the gaze tracking-based method performed better than the traditional method of using only facial expression recognition (Kafi et al., 2019).

Some researchers used body language analysis to detect engagement levels. Studies have used this approach to analyze engagement in online and in-person classes, and have shown that it can be effective in detecting engagement levels. This can include analyzing the position and movement of a student's body, such as sitting or standing, as well as the position of their head and gaze (İbrahimoğlu, 2018). Research in this area has shown that body posture analysis can be used to detect engagement levels. The authors (Bhardwaj et al., 2021) used a dataset of 800 students and achieved an accuracy rate of 89.3% in detecting engagement levels. Many researchers used keystroke dynamics, which is the study of typing patterns, such as typing speed and errors, to detect engagement levels. Research has shown that keystroke dynamics can be used to detect engagement levels and can be used to improve the effectiveness of teaching.

Another approach used head pose and body movements to detect engagement levels. This can include analyzing the position of the head, eye gaze, and body posture. Research in this area has shown that head pose and body movements can be used to detect engagement levels and can be used to enhance the quality of teaching.

2 Related Work

Research has demonstrated that facial expressions serve as a dependable measure of student engagement, and monitoring these expressions can offer meaningful insights into students' emotional states (Gupta et al., 2016). In a study (Vaswani, 2017) students who were more engaged in a class had more positive facial expressions, such as smiles and nods, than students who were less engaged. Similarly, (Abedi&

Khan, 2021) students who were more confused had more negative facial expressions, such as furrowed brows or confusion, than students who were not confused. If the learning technique is effective, the method may achieve better performance even with different challenges like occlusion and pose variations (Whitehill et al., 2014; Jiang & Jia, 2011). Studies (Fakhar et al., 2022) were also done to observe the usage of signals to judge customer engagement based on facial expressions.

Another study (Escobedo et al., 2024; Bidwell & Fuchs, 2011) used an approach to detect the level of engagement using a combination of head pose, facial expression, and gaze analysis (Gao et al., 2021). The suggested approach was evaluated in a real classroom environment and showed promising results (Huang et al., 2019; Mao et al., 2009) proposed a deep learning method for recognizing student engagement in video-based online classes by using facial expression, head pose, and gaze features. They found that their method was able to achieve high accuracy in detecting engagement levels (Islam & Hossain, 2021; De Carolis et al., 2019) retrieved the facial feature by Histogram of Oriented Gradients which is further learned by CNN. This research work gave better performance since the Histogram of Oriented Gradient extracts the spatial oriented gradients' information (Mitra & Acharya, 2024).

Though many worked on facial expression-based engagement analysis, challenges like occlusion and pose variation lower the system's performance. Researchers (Yovel & Duchaine, 2006; Gupta et al., 2016; Gupta et al., 2019) have made some efforts by utilizing texture features or reconstructed geometric features for addressing occlusion. An enhanced Kanade-Lucas tracker (Yovel & Duchaine, 2006) is proposed for recovering drifted and lost facial points. Reconstruction of missing points using PCA-based techniques (Wang et al., 2018). Another initiative named the modified transferable belief model (Gupta et al., 2019) was proposed for recognizing expressions of face. The major contributions of this work are

- A novel framework named Bi-upsampling Feature Pyramid Network (BiusFPN) is proposed using ResNet18 as a backbone network.
- Multiple attention map integration is incorporated along with the proposed inter-coordinate attention model.
- Extensive results and discussion are done with state-of-art datasets.

3 Methodology

This study explores the use of facial expression recognition technology to assess class engagement. Facial expression-based engagement of a person during online mode is identified with the help Feature Pyramid Network (FPN) with a specialized attention mechanism. This approach helps to identify specific facial expressions that are associated with different levels of engagement in an effective way.

3.1. Proposed Inter Cross Coordinate Self Attention

The attention concept plays a vital role in all computer vision areas, in this work a new form of Inter Cross Coordinate Self Attention (ICCSA) is designed to handle the various levels of facial features to clearly distinguish the state of presence. This attention mechanism is the combination of StandAlone Self Attention and Co-ordinate Attention to extract local features in two different variant sampling. The following section explains coordinated attention and how it is inter-crossed with stand-alone self-attention.

3.1.1. Cross Coordinate Attention

Coordination Attention (Ma et al., 2021) compresses global spatial features into a channel descriptor. While global pooling is commonly used in channel attention to encode spatial information at a global level, it often struggles to preserve positional data, which is essential for capturing structures in vision applications.

The author (Ma et al., 2021) factorizes the global pooling designed to induce attention blocks to record lengthy interactions spatially with accurate positioning information as,

$$z_{ch} = \frac{1}{Height* \ Width} \sum_{i=1}^{Height} \sum_{j=1}^{Width} k_{ch}(i,j) \quad (1)$$

into two 1D feature encoding procedures. In particular, given the input K, every channel is encoded along horizontal coordinates and then the vertical coordinates based encoding respectively, using 2 spatial dimensions of pooling kernels such as *(Height, 1), (1, Width).* Thus, the channel output ch-th's at height *h* can be written as equation 1.

$$z_{ch}^{h}(h) = \frac{1}{\text{Width}} \sum_{0 \le i < \text{Width}} k_{ch}(h,i) \quad (2)$$

Similar to that, the *ch*-th channel's output at width *wi* can be expressed as

$$z_{ch}^{w}(\text{width}) = \frac{1}{\text{Height}} \sum_{0 \le j < \text{Height}} k_{ch}(j,w) \quad (3)$$

As a consequence of both of the transformations discussed above, attributes are combined in line with the 2 spatial patterns to create 2 feature maps that follow instructions into consideration as shown in Figure 1.

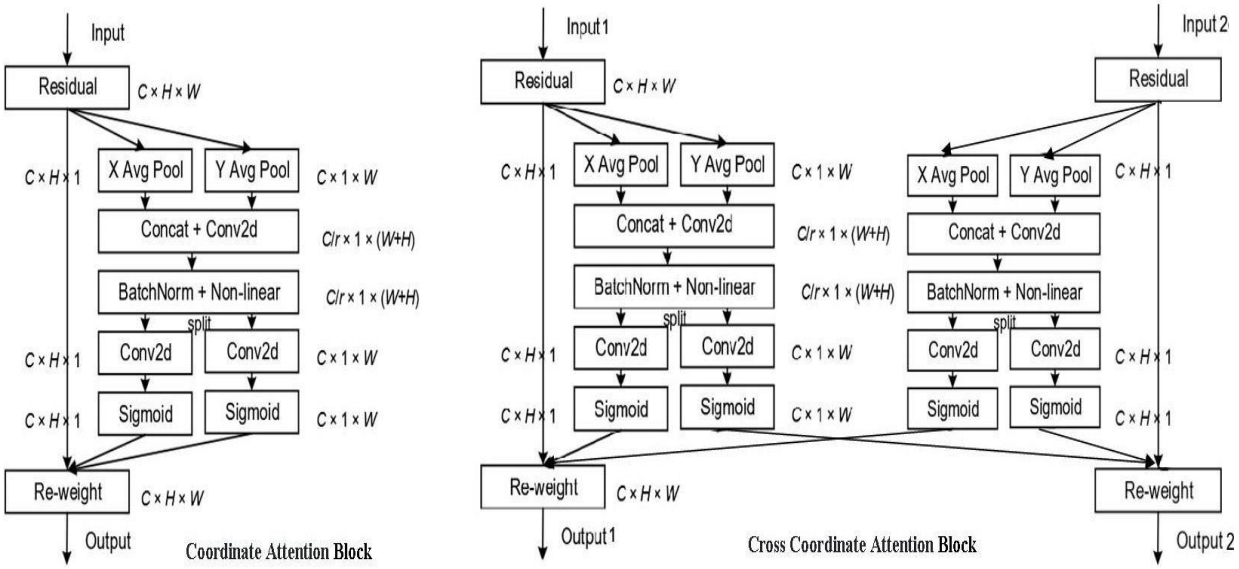

Figure 1: Coordinate Attention and Cross-Coordinate Attention Block

Equations (2) and (3), as noted before, permit a global receptive field and encode precise positional information. The strength of Coordination attention is collecting the positional data to clearly represent the region of interest as well as identification of channel relationships which have been shown to be crucial in previous studies (Gupta et al., 2016; Batra et al., 2022). To be more precise, the model

concatenates the aggregated feature maps obtained Prior to applying Equations 4 and 5 to the typical 1×1 convolution function $\boldsymbol{Func_1}$, which yields

$$f = \delta(Func1([z^h, z^w])) \quad (4)$$

Where [·,·] signifies the concatenation function all across spatial dimension, a non-linear activation function δ and $f \in \mathbb{R}^{C/re\ X\ (Height+Width)}$ is the intermediate feature map that encodes spatial data in both horizontal and vertical directions. The reduction ratio (*re)* is the same used in the SE block to limit block dimension. Then, we divided *f* into 2 distinct tensors across the spatial dimension (Mohammed & Jahanbakhshian, 2021).

$$f^h \in \mathbb{R}^{C/rxHeight} and\ f^w \in \mathbb{R}^{C/rxWidth} \quad (5)$$

A second 1×1 convolutional transformation individually $Func_h$ and $Func_w$ are transformed using f^h and f^w into tensors that have the identical channel number as the input X, resulting in equation 6 & 7.

$$out^h = \sigma(Func_h(f^h)) \quad (6)$$

$$out^w = \sigma(Func_w(f^w)) \quad (7)$$

The outcomes out^h and out^w are then enlarged and applied as the appropriate attention weights. As shown in the Figure 1 the Cross Coordinate Attention model, the concept of coordinate attention is applied for two inputs as it is then it will produce attention weights in horizontal wise as well as vertical wise for two inputs (k_{ch1} and k_{ch2}) which is represented as out_1^h, out_1^w *and* out_2^h, out_2^w with the help of the equation 7 individually for each input. Then the final outcome of the block Y_1 and Y_2 for two inputs such as k_{ch1} and k_{ch2} is derived as, equation 8.

$$y_1(i,j) = k_{ch1}(i,j)\ x\ out_1^h(i)\ x\ out_2^w(j)$$

$$y_2(i,j) = k_{ch2}(i,j)\ x\ out_2^h(i)\ x\ out_1^w(j) \quad (8)$$

This proposed cross coordinate attention generates the feature map from two different coordination-based attention co-efficients. This concept is added the strength of feature extraction by merging two feature maps to represent the expression-oriented characteristics in an efficient manner.

3.1.2. Inter Cross Coordinate Self Attention

A stand-alone self-attention (Ramachandran et al., 2019) can be employed to replace spatial convolutions and construct a fully attention-based model. Similar to a convolution, for a given pixel $x_{ij} \in \mathbb{R}^{d_{in}}$, it first extracts a local region of pixels in positions $ab \in \mathcal{N}_k(i,j)$ with spatial extent k centered around x_{ij} , which we call the memory block. This approach to local attention differs from previous work in vision that has explored global attention (i.e., all-to-all interactions between all pixels) (Wang et al., 2018; Bello et al., 2019). Global attention is computationally intensive and can only be applied after significant spatial downsampling of the input, which limits its use across all layers in a fully attention-based model.

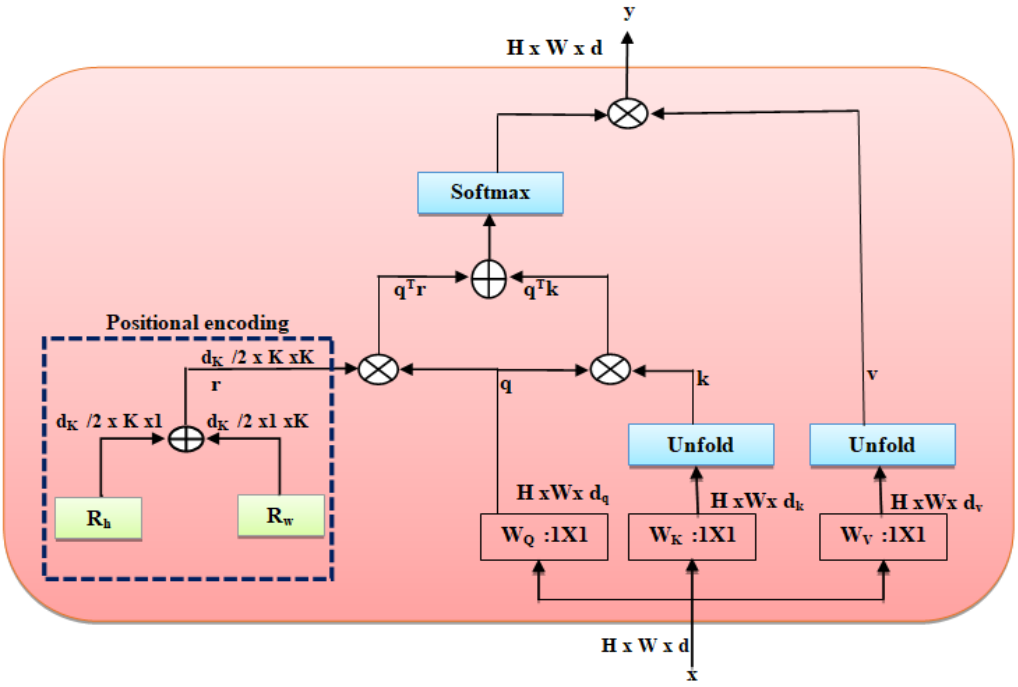

Figure 2: Stand-Alone Self Attention

Single-headed attention for computing the pixel output $y_{ij} \in \mathbb{R}^{d_{out}}$ is calculated as follows in equation 9:

$$y_{ij} = \sum_{a,b \in N_k(i,j)} \text{Softmax}_{ab}(q_i T_j k_{ab}) v_{ab} \quad (9)$$

Here, the queries $q_{ij} = W_Q x_{ij}$, keys $k_{ab} = W_K x_{ab}$, and values $v_{ab} = W_v x_{ab}$ are linear transformations of the pixel at position ij and its neighborhood pixels. $Softmax_{ab}$ denotes a softmax function applied to all logits computed within the neighbourhood of ij. $W_Q, W_K, W_V \in \mathbb{R}^{d_{out} x d_{in}}$ are learned transforms. This local self-attention mechanism aggregates spatial information over neighbourhoods similar to convolutions. But this attention lack in positional information hence the (Ramachandran et al., 2019) Stand alone self attention is proposed to handle this problem with the help of relative attention concept instead of relative positional embeddings (Vaswani, 2017) which improves the performance. In the SASA model along with the single-headed concept, the relative attention concept is also added to produce the final attention map which is shown in Figure 2. Hence the equation (10) becomes like this

$$y_{ij} = \sum_{a,b \in N_k(i,j)} \text{Softmax}_{ab}(q_i T_j k_{ab} + q_i T_j r_{a-i,b-j}) v_{ab} \quad (10)$$

Where $a \; and \; b \in \mathcal{N}_k(i,j)$ that neighbor of (i, j) at a particular distance from that it produces two distances: a row offset a − i and column offset b − j. The row and column offsets are associated with an embedding r_{a-i} and r_{b-j} respectively.

The proposed Inter Cross Coordinate Self Attention (ICCSA) includes Coordinate attention which combines the query q and key k and produces two coordinate attention feature maps which are shown in the figure 3.

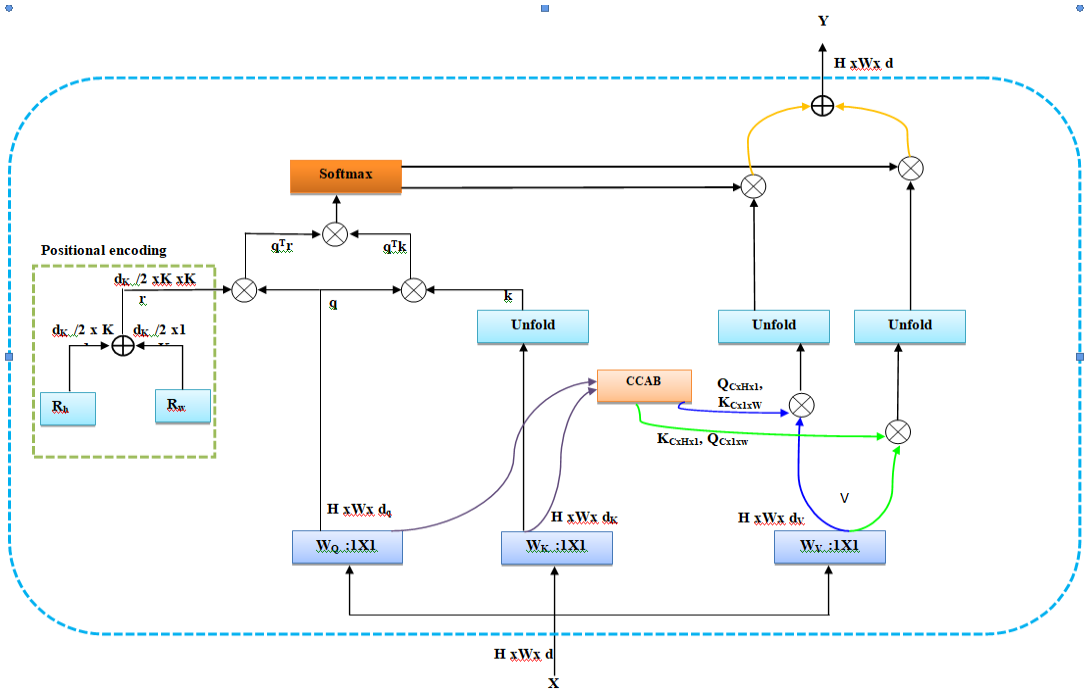

Figure 3: Inter Cross Coordinate Self Attention

For an input with $x_{ij} \in \mathbb{R}^{d_{in}}$ first, the query, key, and values are generated with the help of point convolution with kernel size as 1 x 1. The $Softmax_{ab}$ is estimated as in Stand-alone self-attention. The q and k are further fed into Cross Coordinate Attention Block (CCAB) which is described in section 3.1.1 Two different attention maps are generated from the CCAB using query and key as $Amap\_q_hk_w$ and $Amap\_k_hq_w$ which is further multiplied with v and multiplied with $Softmax_{ab}$ illustreated in the following equations (11 -13).

$$Amap\text{-}q_hk_w \, , \, Amap\text{-}k_hq_w = CCAB(q, k) \tag{11}$$

$$y_{ij}^{1} = \sum_{a,b \in N_k(i,j)} \mathrm{Softmax}_{ab}\left(q_i T_j k_{ab} + q_i T_j r_{a-i,b-j}\right)(v_{ab} * Amap\text{-}q_hk_w) \tag{12}$$

$$y_{ij}^{21} = \sum_{a,b \in N_k(i,j)} \mathrm{Softmax}_{ab}\left(q_i T_j k_{ab} + q_i T_j r_{a-i,b-j}\right)(v_{ab} * Amap\text{-}k_hq_w) \tag{13}$$

The Inter Cross Coordinate Self Attention finally produces the output Y as the summation of y_{ij}^{1} and y_{ij}^{2}.

3.2. Proposed BiusFPN with Inter Cross Coordinate Self Attention Model for Engagement Analysis

The proposed framework for person engagement analysis is designed with the help of Feature Pyramid Network with Resnet-18 as the backbone. This Bi upsampling feature pyramid network handles two times of up sampling for a single flow-down sampling process.

In this model, an Inter Cross Coordinate Self Attention model is introduced along with the traditional channel and spatial attention in the second level of the upsampling process. The overall architecture of the proposed framework is shown in the following figure 4.

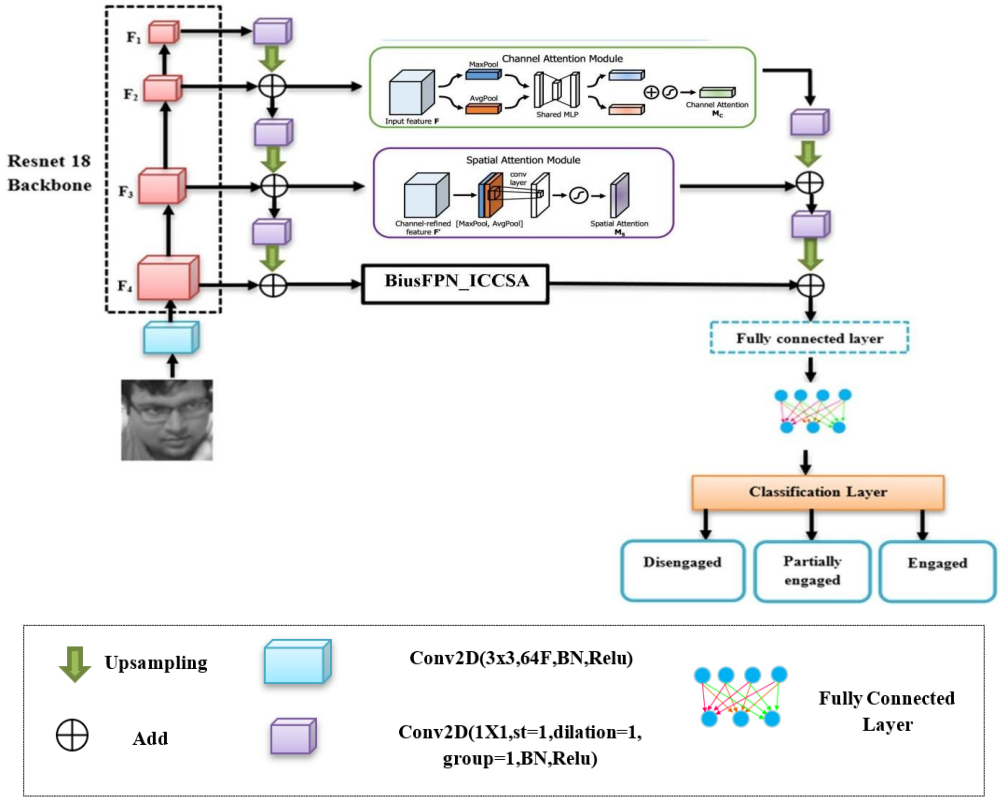

Figure 4: Architecture of the Proposed BiusFPN with Inter Cross Coordinate Self Attention Model

3.2.1. Resnet 18

ResNet18 is an 18-layer Convolutional Neural Network designed for efficient deep learning operations. As networks grow deeper, they face the vanishing gradient problem, where gradients diminish due to repeated multiplications, leading to saturation or performance loss. To address this, ResNet introduces skip connections that bypass certain layers, allowing the model to learn residual mappings instead of direct transformations. This approach improves training stability and prevents gradient-related issues in deep networks. ResNet18 serves as the backbone network for the proposed model. Figure 5 shows the skip connection process of the residual block.

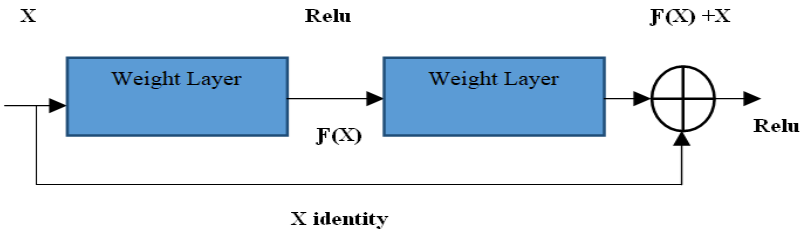

Figure 5: Skip Connections

3.2.2. Feature Pyramid

The Feature Pyramid Network (FPN) is a cost-effective, multi-scale deep learning framework used for object recognition across various scales. It combines high-level semantic features with detailed localization information by integrating features from different layers through top-down and bottom-up

pathways. FPN outperforms traditional methods like Faster R-CNN by generating higher-quality feature maps and is computationally efficient, as it leverages multi-scale feature maps computed during the forward pass. Due to its accuracy, speed, and benefits, FPN is recommended for use in the proposed model with ResNet-18 as the backbone.

Figure 4 illustrates the architecture of the proposed technique. The model takes a 100×100 grayscale facial image as input. The convolutional layer initially processes it with 64 filters (3×3 kernel), producing a 100×100×64 feature map, which is then passed to ResNet18. ResNet18 applies four convolutional layers, generating feature maps:

F4 (100×100×64) → 64 filters, 3×3 kernel, stride 1

F3 (50×50×128) → 128 filters, 3×3 kernel, stride 2

F2 (25×25×256) → 256 filters, 3×3 kernel, stride 2

F1 (13×13×512) → 512 filters, 3×3 kernel, stride 2

Next, Feature Pyramid Network (FPN) upsampling is applied. Feature maps are progressively upsampled and added:

F1 (13×13×512) → add1 (25×25×256) → F2

add1 → add2 (50×50×128) → F3

add2 → add3 (100×100×64) → F4

Finally, the first upsampling output is processed using three attention mechanisms: Channel, Spatial, and ICCSA, before entering the second upsampling stage.

3.2.3. Channel and Spatial Attention

In convolutional neural networks, a channel attention module (Sandler et al., 2018) is used to focus on "what" is important in an input image by leveraging inter-channel feature relationships. It generates a channel attention map $M_c \in \mathbb{R}^{cX1X1}$ by aggregating spatial information using average-pooling and max-pooling to create descriptors F_{avg}^{c} and F_{max}^{C}. These descriptors are processed through a shared multi-layer perceptron (MLP) with a hidden layer of size $\mathbb{R}^{c/rX1X1}$ to reduce parameters, where r is the reduction ratio. The outputs are combined via element-wise summation to produce the final channel attention map. The channel attention is calculated as follows, in brief in equation 13 & 14:

$$M_c(F) = \sigma(MLP(AvgPool(F) + MLP(MaxPool(F))) \quad (13)$$

$$M_c(F) = \sigma(W_1\left(W_0(F_{avg}^{c})\right) + W_1(W_0(F_{max}^{C}))) \quad (14)$$

The sigmoid function (σ) is applied where$W_0 \in \mathbb{R}^{c/rXc}$, and$W_1 \in \mathbb{R}^{cXc/r}$, with MLP weights following the ReLU activation for both inputs. Figure 6 illustrates the channel attention architecture, which focuses on identifying important features.

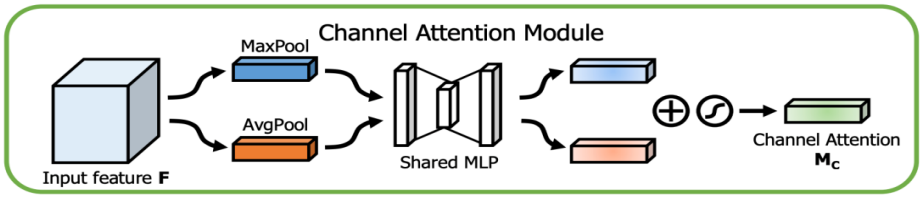

Figure 6: Channel Attention

Spatial attention, a module in CNNs, emphasizes where information is relevant by leveraging spatial relationships between features. Unlike channel attention, which focuses on what is important, spatial attention uses average-pooling and max-pooling along the channel axis to generate an effective feature descriptor. A convolution layer then processes this descriptor to create a spatial attention $M_s(F) \in \mathbb{R}^{HXW}$, highlighting areas for emphasis or suppression. Two 2D maps, $F^s_{avg} \in \mathbb{R}^{1XHXW}$ and $F^s_{max} \in \mathbb{R}^{1XHXW}$, representing average and max-pooled features, are concatenated and convolved to generate the final spatial attention map. The computation of the spatial attention is, in brief in equation 15 & 16:

$$M_s(F) = \sigma\left(f^{7x7}([AvgPool(F); MaxPool(F)])\right) \quad (15)$$

$$M_s(F) = \sigma\left(f^{7x7}\left(\left[F^s_{avg}; F^s_{max}\right]\right)\right) \quad (16)$$

The sigmoid function (σ) and f^{7x7} represent a convolution operation with a filter size of 7 × 7. The architecture of Spatial Attention is depicted in Figure 7.

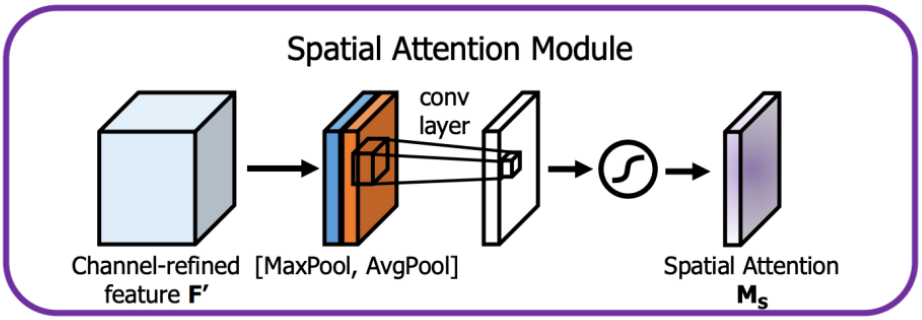

Figure 7: Spatial Attention

The feature maps obtained in the first level of upsampling will be fed into our proposed BiusFPN ICCSA and also to channel attention and spatial attention. *Add1* will be used as an input for channel attention. The output will be the same size as the input (25×25×256). The convolutional layer having filter size of 1×1, stride 1, dilation 1, group 1, Batch Normalization, and Relu with a number of filters of 128 receives the output of the channel attention. It is then added to the output of spatial attention after being routed into up-sampling. Addition *add4* as a consequence will be 50×50×128. This will once more be fed into a convolution layer, but this time, the filter size will be 64. It is then up-sampled and added to the results of our suggested BiusFPN ICCSA. The resulting addition *add5* will therefore be 100×100×64. This will then be delivered into a fully connected layer with a filter size of 1086 and a final fully connected layer with four classes as the classification layer will produce results such as Disengaged, Partially engaged, or Engaged after the fully connected layer, accurately classifying the input image.

4 Results and Discussion

4.1. Dataset Description

4.1.1. Daisee Dataset

The DAiSEE dataset (Liao et al., 2021) consists of 9,068 video clips from 112 students in an e-learning environment, annotated for engagement, frustration, confusion, and boredom on a 4-level scale (0: very low to 3: very high). The videos are 10 seconds long, 30 fps, with a 640×480 resolution. This study focuses only on engagement classification.

4.1.2. WACV Dataset

The WACV dataset (Ling et al., 2012) contains 4,424 images categorized into disengaged, moderately engaged, and engaged classes. The dataset is imbalanced, with 412 disengaged, 2,247 moderately engaged, and 1,765 engaged images. To balance it, 412 images were randomly selected from each class, reshaping them to 100×100×3 for uniformity.

We divided this data into training and testing (80% and 20%, respectively). The graphics below show examples from the WACV dataset and the DAiSEE using class-wise.

Figure 8: Engaged (top), Partially Engaged (middle), and Disengaged (bottom) Samples of the WACV Dataset

Figure 8 shows the Disengaged, Partially Engaged, and Engaged samples of the WACV dataset, and Figure 9 shows the Boredom, Confusion, Frustration, and Engagement samples of the DAiSEE dataset.

Figure 9: Engagement (first row), Boredom (second row), Confusion (third row), and Frustration (bottom row) Samples of the DAiSEE Dataset

The proposed model BiusFPN_ICCSA yielded a higher accuracy of 68.16% compared with other existing methods. Table 1 shows the accuracy of different methods for the DAiSEE dataset.

Table 1: Accuracy Comparison of DAiSEE Dataset for Different Methods

Methods	Accuracy (%)
C3D (Gupta et al., 2016)	48.1
I3D (Zhang et al., 2019)	52.4
C3D + LSTM (Abedi & Khan, 2021)	56.6
C3D with transfer learning (Gupta et al., 2016)	57.8
LRCN (Gupta et al., 2016)	57.9
DFSTN (Liao et al., 2021)	58.8
C3D + TCN (Abedi & Khan, 2021)	59.9
ResNet + LSTM (Abedi & Khan, 2021)	61.5
ResNet + TCN (Abedi & Khan, 2021)	63.9
DERN (Huang et al., 2019)	60
Neural Turing Machine (Ma et al., 2021)	61.3
latent affective + behavioral + affect (TCN model) (Abedi & Khan, 2021)	63.3
ResNet-18 (Batra et al., 2022)	66.64
BiusFPN_ICCSA	**68.16**

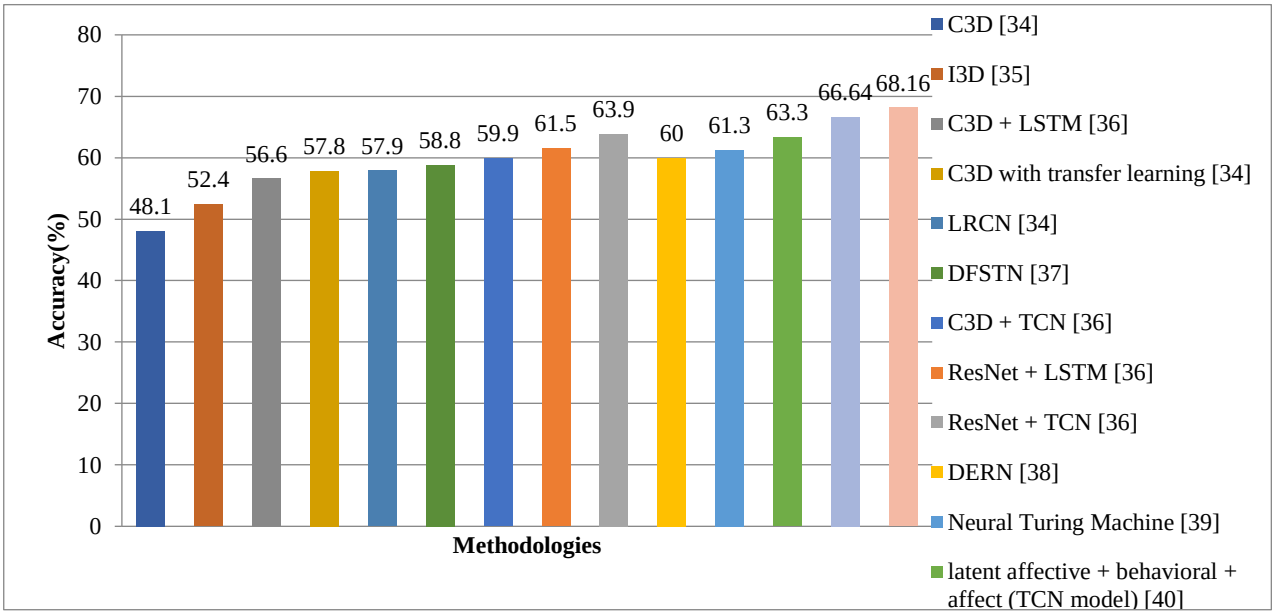

Figure 10: Accuracy Comparison of DAiSEE Dataset with Existing Methods

The graphical representation of the accuracy comparison of DAiSEE dataset with Existing Methods is given in Figure 10. In terms of Accuracy, BiusFPN_ICCSA gives +3.5 greater than latent affective (TCN model) (Abedi & Khan, 2021), +5.5 greater than Neural Turing Machine (Ma et al., 2021), +6.8 greater than DERN (Huang et al., 2019), +2.9 greater than ResNet+TCN (Abedi & Khan, 2021), +5.3 greater than ResNet+LSTM (Abedi & Khan, 2021). +6.9 greater than C3D+TCN (Abedi & Khan, 2021), +8 greater than DFSTN (Liao et al., 2021), +8.9 greater than LRCN (Gupta et al., 2016), +9 greater than C3D with transfer learning (Gupta et al., 2016), +10.2 greater than C3D+LSTM (Abedi & Khan, 2021), +14.4 greater than 13D (Zhang et al., 2019), and +18.7 greater than C3D (Gupta et al., 2016). Table 2 shows the Accuracy, AUC, Gini index, and AGF comparison of Resnet-18 (Batra et al., 2022) and BiusFPN_ICCSA methods for classes such as Boredom, Confusion, Frustration, and Engagement.

Table 2: Metrics Comparison of Resnet 18 and BiusFPN_ICCSA Methods

Classes	ResNet18 (Batra et al., 2022)				BiusFPN_ICCSA			
	Accuracy	AUC	Gini Index	AGF	Accuracy	AUC	Gini Index	AGF
Boredom	75	86.99	73.98	63.59	75	87.19	74.38	70.65
Confusion	80.95	88.09	76.18	82.98	84.52	90.14	80.28	85.44
Frustration	61.45	69.69	39.38	66.31	62.58	70.53	41.07	67.25
Engagement	70.76	70.07	40.14	71.32	72.48	71.24	42.48	72.68

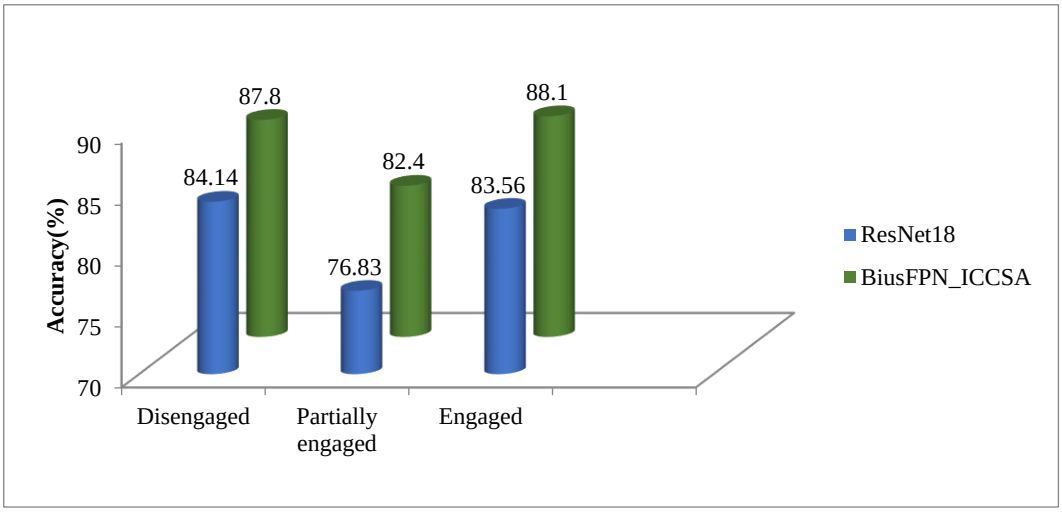

Figure 11: Class-wise Comparison of the Accuracy of Resnet18 and BiusFPN_ICCSA for the DAiSEE Dataset

Figure 11 shows the Class-wise comparison of accuracy for Resnet18 and BiusFPN_ICCSA methods. In terms of accuracy, ResNet and the proposed BiusFPN_ICCSA are similar for boredom class. The proposed BiusFPN_ICCSA method is +3.57 greater than ResNet for the confusion class, +1.13 greater for the frustration class, and +1.72 greater for the engagement class.

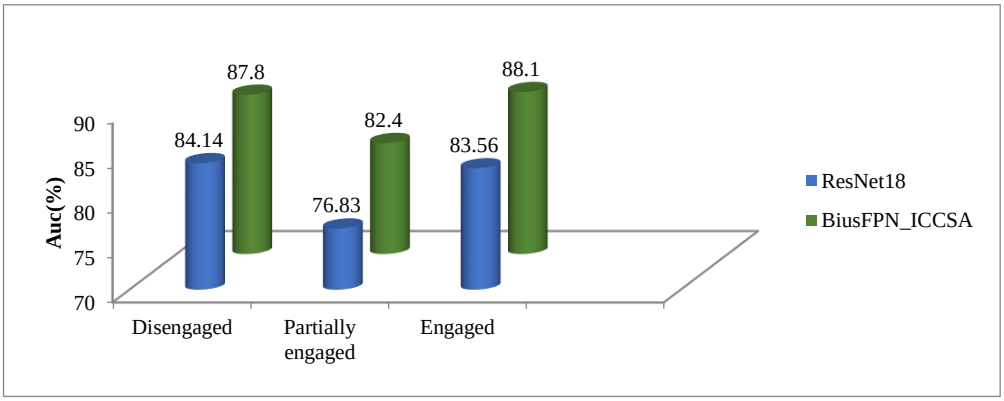

Figure 12: Classwise Comparison of AUC of Resnet18 and BiusFPN_ICCSA for DAiSEE Dataset

Figure 12 shows the Class-wise comparison of AUC for Resnet18 and BiusFPN_ICCSA methods. In terms of AUC, the proposed BiusFPN_ICCSA is +0.2 greater than ResNet for the boredom class, +2.05 greater than for confusion class, +0.84 greater than for frustration class, and +0.48 greater for the engagement class.

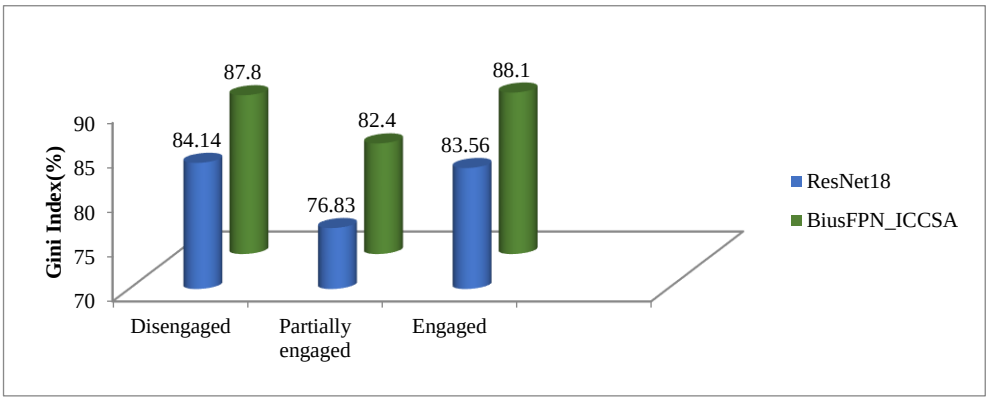

Figure 13: Classwise Comparison of Gini Index of Resnet18 and BiusFPN_ICCSA for the DAiSEE Dataset

Figure 13 shows the Class-wise comparison of the Gini Index for Resnet18 and BiusFPN_ICCSA methods. In terms of the Gini Index, the proposed BiusFPN_ICCSA is +0.4 greater than ResNet for the boredom class, +4.1 greater than for the confusion class, +1.69 greater than for the frustration class, and +2.34 greater for the engagement class.

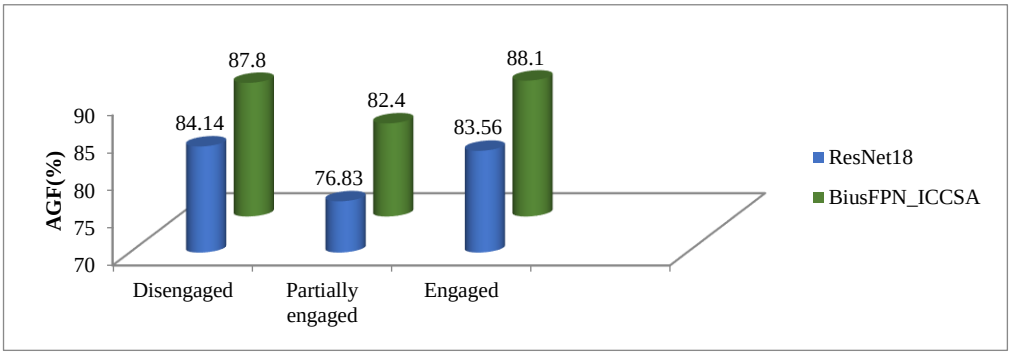

Figure 14: Classwise Comparison of AGF of Resnet18 and BiusFPN_ICCSA for the DAiSEE Dataset

Figure 14 shows the Class-wise comparison of AGF for Resnet18 and BiusFPN_ICCSA methods. In terms of the Gini Index, the proposed BiusFPN_ICCSA is +7.06 greater than ResNet for the boredom class, +2.46 greater than for the confusion class, +0.94 greater than for frustration class, and +1.36 greater for engagement class.

Table 3 shows the accuracy of different methods for the WACV dataset.

Table 3: Accuracy of Different Methods for the WACV Dataset

Methods	Accuracy
CNN (Islam & Hossain, 2021)	37
HOG+SVM (Batra et al., 2022)	58
HOG+CNN (Batra et al., 2022)	42
HOG+SIFT+SVM (Batra et al., 2022)	32
SURF+SVM (Batra et al., 2022)	35
DenseNet-121 (Batra et al., 2022)	78
ResNet-18 (Batra et al., 2022)	80
MobileNetV1 (Batra et al., 2022)	66
BiusFPN with Inter Cross Coordinate Self Attention	**83**

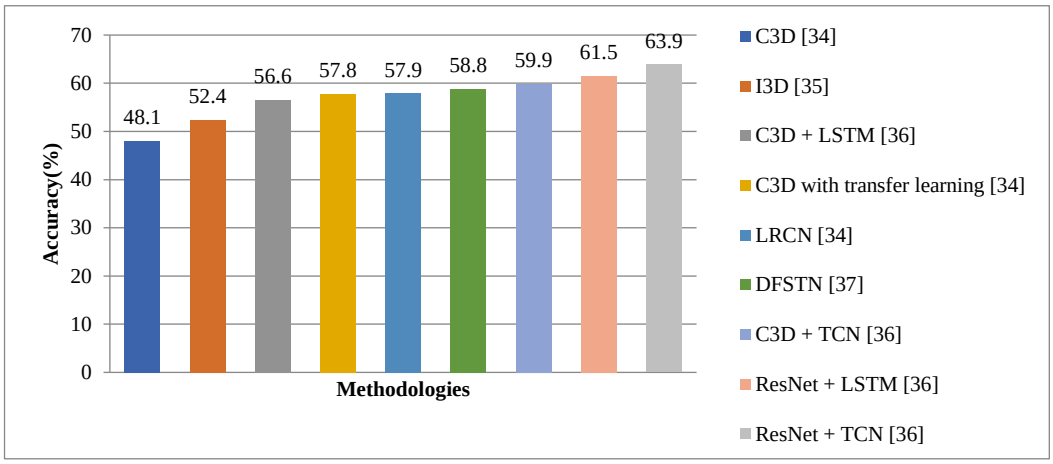

Figure 15: Accuracy Comparison of WACV Dataset with Existing Methods

The graphical representation of the accuracy comparison of the WACV dataset with Existing Methods is given in Figure 15. In terms of Accuracy, proposed method gives +17 greater than MobileNetV1 (Ordoubadi, 2017), +3 greater than ResNet-18 (Ordoubadi, 2017), +5 greater than DenseNet-121 (Batra et al., 2022), +48 greater than SURF+SVM (Ordoubadi, 2017), +51 greater than HOG+SIFT+SVM (Ordoubadi, 2017), +41 greater than HOG+CNN (Batra et al., 2022), +25 greater than HOG+SVM(Ordoubadi, 2017), and +46 greater than CNN (Pabba & Kumar, 2022).

Table 4: Metrics Comparison of Resnet 18 and BiusFPN_ICCSA Methods

Classes	ResNet18				BiusFPN_ICCSA			
	Accuracy	AUC	Gini Index	AGF	Accuracy	AUC	Gini Index	AGF
Disengaged	84.14	90.03	80.07	87.08	87.8	90.22	80.44	87.4
Partially engaged	76.83	84.3	68.6	83.31	82.4	88.32	76.65	85.08
Engaged	83.56	90.09	80.18	88.49	88.1	89.62	79.25	89.82

Figure 16 & table 4 shows the Class-wise comparison of Accuracy for Resnet18 and BiusFPN_ICCSA methods. In terms of accuracy, the proposed BiusFPN_ICCSA is +3.66 greater than ResNet for the disengaged class, +5.57 greater than for the partially engaged class, and +4.85 greater than for the engaged class.

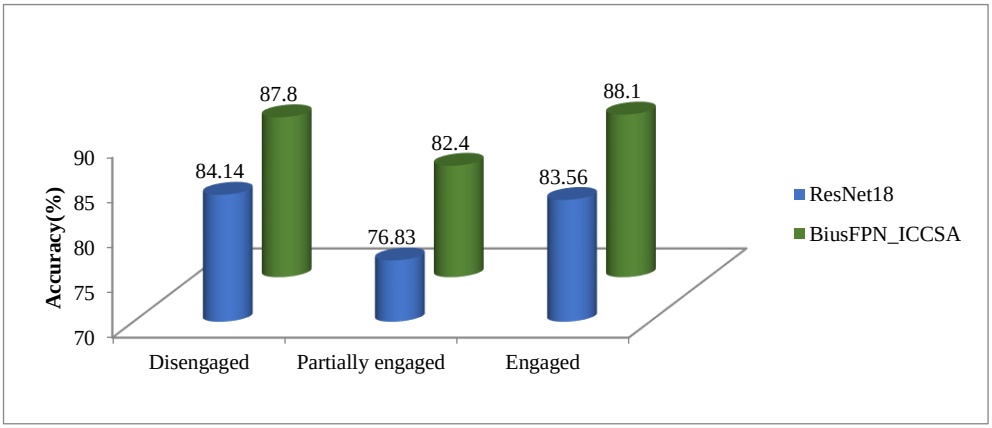

Figure 16: Class-wise Comparison of Accuracy of Resnet18 and BiusFPN_ICCSA for the WACV Dataset

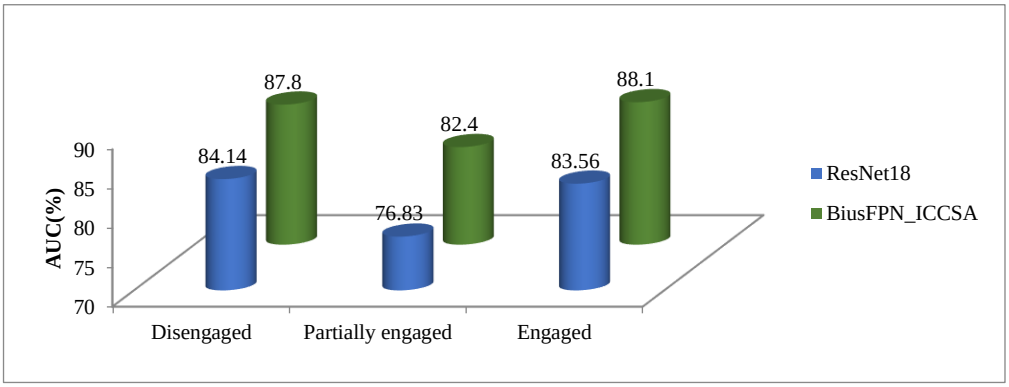

Figure 17: Classwise Comparison of AUC of Resnet18 and BiusFPN_ICCSA for the WACV Dataset

Figure 17 shows the Class-wise comparison of AUC for Resnet18 and BiusFPN_ICCSA methods. In terms of AUC, the proposed BiusFPN_ICCSA is +0.19 greater than ResNet for the disengaged class, +4.02 greater than for the partially engaged class, and -0.47 lesser than for the engaged class. Figure 18 shows the Class-wise comparison of the Gini Index for Resnet18 and BiusFPN_ICCSA methods. In terms of the Gini Index, the proposed BiusFPN_ICCSA is +0.37 greater than ResNet for the disengaged class, +8.05 greater than for the partially engaged class, and -0.93 lesser than for the engaged class.

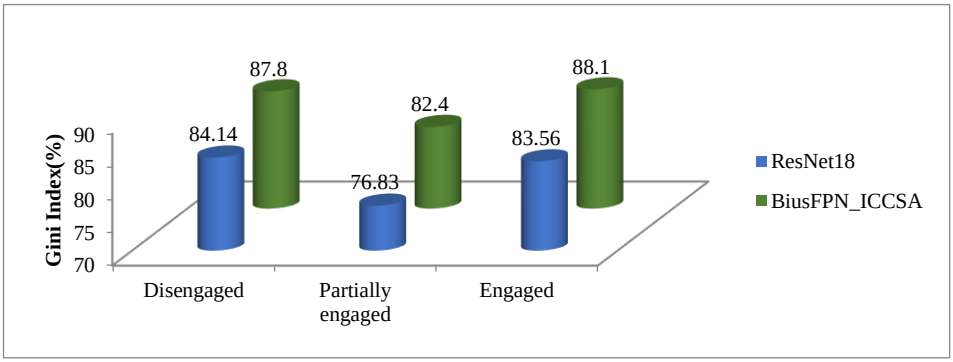

Figure 18: Classwise Comparison of Gini Index of Resnet18 and BiusFPN_ICCSA for the WACV Dataset

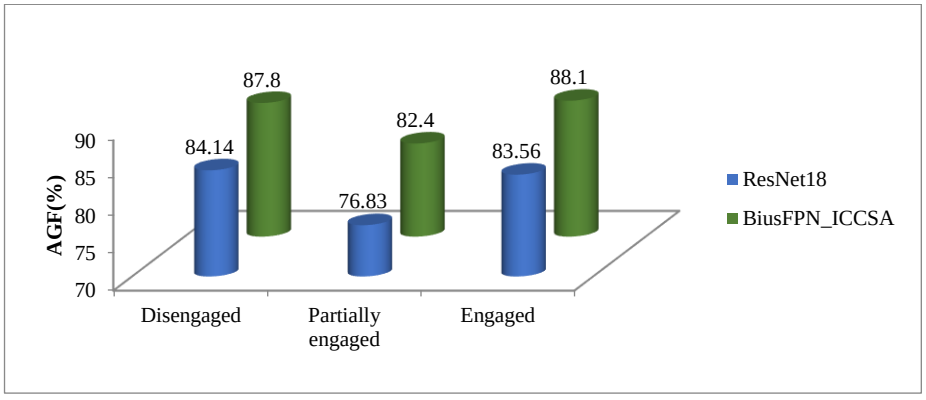

Figure 19: Classwise Comparison of AGF of Resnet18 and Biusfpn_Iccsa for The WACV Dataset

Figure 19 shows the Class-wise comparison of AGF for Resnet18 and BiusFPN_ICCSA methods. In terms of AGF, the proposed BiusFPN_ICCSA is +0.32 greater than ResNet for the disengaged class, +1.77 greater than for the partially engaged class, and +1.33 greater than for the engaged class.

5 Conclusion

Facial expression recognition technology in classrooms has the potential to improve learning outcomes and provide valuable insights into students' emotional states. The results of this study will deepen our understanding of the relationship between facial expressions and engagement, guiding the development of more effective strategies for tracking and enhancing student participation in class. In this work, the Feature Pyramid model is enhanced to attain various scales of features which is further analyzed to carry over the most related features for different classes with the help of various attention techniques. The relative attention model of Stand-alone self-attention is integrated with coordinate attention in the form of a cross-oriented feature map joining to obtain different orientation-based features to effectively distinguish various expressions. The proposed approach Bi-upsampling FPN with the newly designed Inter Cross Coordinate Self Attention mechanism achieves improved performance compared to the stat art of works.

References

[1] Abedi, A., & Khan, S. (2021). Affect-driven ordinal engagement measurement from video. https://doi.org/10.48550/arXiv.2106.10882

[2] Abedi, A., & Khan, S. S. (2021, May). Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. In *2021 18th Conference on Robots and Vision (CRV)* (pp. 151-157). IEEE. https://doi.org/10.1109/CRV52889.2021.00028

[3] Amraee, M., & Koochari, A. (2014). Face recognition using a training sample from each individual. *International Academic Journal of Innovative Research*, *1*(2), 6–13.

[4] Batra, S., Wang, H., Nag, A., Brodeur, P., Checkley, M., Klinkert, A., & Dev, S. (2022). DMCNet: Diversified model combination network for understanding engagement from video screengrabs. *Systems and Soft Computing*, *4*, 200039. https://doi.org/10.1016/j.sasc.2022.200039

[5] Bello, I., Zoph, B., Vaswani, A., Shlens, J., & Le, Q. V. (2019). Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3286-3295).

[6] Bhardwaj, P., Gupta, P. K., Panwar, H., Siddiqui, M. K., Morales-Menendez, R., & Bhaik, A. (2021). Application of deep learning on student engagement in e-learning environments. *Computers & Electrical Engineering*, *93*, 107277. https://doi.org/10.1016/j.compeleceng.2021.107277

[7] Bidwell, J., & Fuchs, H. (2011). Classroom analytics: Measuring student engagement with automated gaze tracking. *Behav Res Methods*, *49*(113).

[8] Bourel, F., Chibelushi, C. C., & Low, A. A. (2001, September). Recognition of Facial Expressions in the Presence of Occlusion. In *BMVC* (pp. 1-10).

[9] De Carolis, B., D'Errico, F., Macchiarulo, N., & Palestra, G. (2019, October). "Engaged Faces": Measuring and Monitoring Student Engagement from Face and Gaze Behavior. In *IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume* (pp. 80-85. https://doi.org/10.1145/3358695.3361748

[10] Escobedo, F., Canales, H. B. G., Criollo, R. A. G., Romero, E. M. Y., Orellana, C. S., Ramírez, J. A. B., Vela, C. A. L., & Herrera, J. M. G. (2024). A Smart Crowd Monitoring and Management Model for Humanity in Intelligent Environments: A Real-Time ApplicationScenario. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 15*(4), 166-178. https://doi.org/10.58346/JOWUA.2024.I4.011

[11] Eskandarian, M., AghaSadeghi, S., Tabasi, S., Zerafatkhah, S., & Sheikhian, M. (2016). An Empirical Study: Effect of Emotional Intelligence Components and Self-Confidence on Social

Capital (Case Study: Teachers of Kashan County). *International Academic Journal of Organizational Behavior and Human Resource Management*, *3*(1), 24–32.

[12] Fakhar, S., Baber, J., Bazai, S. U., Marjan, S., Jasinski, M., Jasinska, E., ... & Hussain, S. (2022). Smart classroom monitoring using novel real-time facial expression recognition system. *Applied Sciences*, *12*(23), 12134. https://doi.org/10.3390/app122312134

[13] Gao, M., Song, P., Wang, F., Liu, J., Mandelis, A., & Qi, D. (2021). A Novel Deep Convolutional Neural Network Based on ResNet-18 and Transfer Learning for Detection of Wood Knot Defects. *Journal of Sensors*, *2021*(1), 4428964. https://doi.org/10.1155/2021/4428964

[14] Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. https://doi.org/10.48550/arXiv.1609.01885

[15] Gupta, R. K., Aggarwal, A., & Tiwari, R. K. (2017). Automated detection of student engagement using head pose and body language. In *2017 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-6).

[16] Gupta, S. K., Ashwin, T. S., & Guddeti, R. M. R. (2019). Students' affective content analysis in smart classroom environment using deep learning techniques. *Multimedia Tools and Applications*, *78*, 25321-25348. https://doi.org/10.1007/s11042-019-7651-z

[17] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[18] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13713-13722).

[19] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).

[20] Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Hawthorne, C., Dai, A. M., Hoffman, M. D., & Eck, D. (2018). Music transformer. In *Advances in Neural Information Processing Systems*.

[21] Huang, T., Mei, Y., Zhang, H., Liu, S., & Yang, H. (2019, July). Fine-grained engagement recognition in online learning environment. In *2019 IEEE 9th international conference on electronics information and emergency communication (ICEIEC)* (pp. 338-341). IEEE. https://doi.org/10.1109/ICEIEC.2019.8784559

[22] İbrahimoğlu, D. (2018). The relationship between spiritual intelligence, gendaruality and happiness among students of Islamic Azad University, District 13. *International Academic Journal of Business Management*, *5*(1), 1–17. https://doi.org/10.9756/IAJBM/V5I1/1810001

[23] Islam, M. S., & Hossain, E. (2021). Foreign exchange currency rate prediction using a GRU-LSTM hybrid network. *Soft Computing Letters*, *3*, 100009. https://doi.org/10.1016/j.socl.2020.100009

[24] Jiang, B., & Jia, K. B. (2011). Research of robust facial expression recognition under facial occlusion condition. In *Active Media Technology: 7th International Conference, AMT 2011, Lanzhou, China, September 7-9, 2011. Proceedings 7* (pp. 92-100). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23620-4_13

[25] Kafi, D. K., Ayyash, A. N., & Mohammed, A. S. (2019, February). Theoretical Study of Structural Properties and Energies of a 2-Aminophenol-Vanillin Molecule. In *Journal of Physics: Conference Series* (Vol. 1178, No. 1, p. 012007). IOP Publishing. https://doi.org/10.1088/1742-6596/1178/1/012007

[26] Kaur, A., Mustafa, A., Mehta, L., & Dhall, A. (2018, December). Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)* (pp. 1-8). IEEE.

[27] Lemay, D. J., Basnet, R. B., & Doleck, T. (2020). Examining the Relationship between Threat and Coping Appraisal in Phishing Detection among College Students. *Journal of Internet Services and Information Security*, *10*(1), 38-49.

[28] Liao, J., Liang, Y., & Pan, J. (2021). Deep facial spatiotemporal network for engagement prediction in online learning. *Applied Intelligence*, *51*(10), 6609-6621. https://doi.org/10.1007/s10489-020-02139-8

[29] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

[30] Ling, A.P.A., Kokichi, S., & Masao, M. (2012). Enhancing Smart Grid System Processes via Philosophy of Security-Case Study based on Information Security Systems-. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, *3*(3), 94-112.

[31] Ma, X., Xu, M., Dong, Y., & Sun, Z. (2021). Automatic student engagement in online learning environment based on neural turing machine. *International Journal of Information and Education Technology*, *11*(3), 107-111.

[32] Mao, X., Xue, Y., Li, Z., Huang, K., & Lv, S. (2009, May). Robust facial expression recognition based on RPCA and AdaBoost. In *2009 10th Workshop on Image Analysis for Multimedia Interactive Services* (pp. 113-116). IEEE. https://doi.org/10.1109/WIAMIS.2009.5031445

[33] Mitra, S., & Acharya, S. C. (2024). Socio-Emotional Well-Being and its Determinants in School Students: A Comprehensive Review. *Indian Journal of Information Sources and Services*, *14*(4), 108–116. https://doi.org/10.51983/ijiss-2024.14.4.17

[34] Mohammed, H. Z., & Jahanbakhshian, P. (2021). Studying the Relationship between Emotional Intelligence and Problem–solving Skills of Individuals from Personal Knowledge Management Perspective. *International Academic Journal of Business Management*, *8*(1), 01–11. https://doi.org/10.9756/IAJBM/V8I1/IAJBM0801

[35] Ordoubadi, S. (2017). The Effectiveness of Cognitive-Behavioral Training on Emotional Self-regulation in Addicts of Urmia. *International Academic Journal of Social Sciences*, *4*(2), 26–32.

[36] Pabba, C., & Kumar, P. (2022). An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition. *Expert Systems*, *39*(1), e12839. https://doi.org/10.1111/exsy.12839

[37] Pravin Kumar, M., Jayaraman, T., Senthilkumar, M., & Sumaiya Begum, A. (2023). Performance Investigation of Generalized Rain Pattern Absorption Attention Network for Single-Image Deraining. *Journal of Circuits, Systems and Computers*, *32*(13), 2350231. https://doi.org/10.1142/S0218126623502316

[38] Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., & Shlens, J. (2019). Stand-alone self-attention in vision models. *Advances in neural information processing systems*, *32*.

[39] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520).

[40] Soozanyar, A., & Jafarzadeh, M. R. (2017). An Investigation of the Status of Student-Student Interaction Element in the Iranian E-Learning System. *International Academic Journal of Humanities*, *4*(2), 150–160.

[41] Tamannaeifar, M., & Hesampour, F. (2016). The Relationship between Cultural and Emotional Intelligence with Students' Adjustment to University. *International Academic Journal of Organizational Behavior and Human Resource Management*, *3*(2), 15–27.

[42] Tang, X. Y., Peng, W. Y., Liu, S. R., & Xiong, J. W. (2020, February). Classroom teaching evaluation based on facial expression recognition. In *Proceedings of the 2020 9th international conference on educational and information technology* (pp. 62-67). https://doi.org/10.1145/3383923.3383949

[43] Uçar, M. U., & Özdemir, E. (2022). Recognizing students and detecting student engagement with real-time image processing. *Electronics*, *11*(9), 1500. https://doi.org/10.3390/electronics11091500

[44] Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

[45] Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7794-7803).

[46] Whitehill, J., Serpell, Z., Lin, Y. C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, *5*(1), 86-98. https://doi.org/10.1109/TAFFC.2014.2316163

[47] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3-19).

[48] Yovel, G., & Duchaine, B. (2006). Specialized face perception mechanisms extract both part and spacing information: Evidence from developmental prosopagnosia. *Journal of Cognitive Neuroscience*, *18*(4), 580-593. https://doi.org/10.1162/jocn.2006.18.4.580

[49] Yu, H., Gupta, A., Lee, W., Arroyo, I., Betke, M., Allesio, D., ... & Woolf, B. P. (2021, July). Measuring and integrating facial expressions and head pose as indicators of engagement and affect in tutoring systems. In *International Conference on Human-Computer Interaction* (pp. 219-233). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-77873-6_16

[50] Zhang, H., Xiao, X., Huang, T., Liu, S., Xia, Y., & Li, J. (2019, July). An novel end-to-end network for automatic student engagement recognition. In *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)* (pp. 342-345). IEEE. https://doi.org/10.1109/ICEIEC.2019.8784507

Authors Biography

Naveen Adusumilli is currently pursuing his Ph.D. at GITAM Deem to be University Bengaluru Campus and working as Assistant Professor in the Department of Computer Science and Engineering at Nalla Narasimha Reddy Education Society's Group of Institutions, Hyderabad, India. He holds a B.Tech and M.Tech in Computer Science and Engineering from Jawaharlal Nehru Technological University Hyderabad (JNTUH).

Dr. Israel Jeena Jacob is an image processing expert. She has completed her B.E in Computer Science and Engineering from St.Xavier's Catholic College of Engineering and M.E in Computer Science and Engineering from Karunya Deemed University. She completed her Ph.D degree in Information and Communication Engineering from Anna University. Her areas of interest are Deep Learning and Computer Vision.

Ajay Kumar Mandava received his B.Tech degree in Electronics and Communications Engineering from Jawaharlal Nehru Technological University, Hyderabad, India in 2006. He received the MS and Ph.D. degree in Electrical Engineering from the University of Nevada, Las Vegas, USA in 2010 and 2013. He is currently an Associate Professor in the Department of Electrical, Electronics & Communication Engineering at GITAM Bengaluru Campus. Prior to joining GITAM, he was a Sr. Software Engineer R & D Consultant at Scientific Games - Las Vegas. His research interests include image processing, computer vision, artificial intelligence, pattern recognition and machine learning, computational intelligence, calculus of variations, partial differential equations, software defined vehicles, nondestructive testing and evaluation.