International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Real Time Emotion Detection in Students with Hybrid Deep Learning Architecture Using Resnet and Zfnet

Ms. K. Thrilochana Devi[1], Uppala Bharath Kumar[2], Bakka Tarun Kumar[3], Ravula Rajeev Chowdary[4], Rachamandi Sai Siva Teja[5]

[1]Assistant Professor
[1, 2, 3, 4, 5]Dept of Information Technology, Vasireddy Venkatadri Institute of Technology Guntur, India

Abstract
The realm of deep learning facilitates the automated extraction of features and the recognition of intricate patterns, rendering it exceptionally proficient for complicated tasks such as image categorization, natural language analysis, and sentiment recognition. Student motivation plays a pivotal role in achieving academic success; however, conventional techniques for evaluating engagement are rooted in subjective assessments. This Paper proposes a real-time sentiment detection system that harnesses a hybrid deep learning structure that fuses ZFNet and ResNet for precise facial expression identification. By utilizing the EmotiW2019 dataset, the model sorts student emotions into engagement-related classifications like "engaged" or "not engaged." The system interprets live webcam input, offering instant visual feedback through emojis and metrics on engagement. By merging cutting-edge computer vision and deep learning methodologies, this adaptable solution amplifies both physical and online learning settings, empowering data-informed interventions to enhance student involvement and educational outcomes. Experimental findings reveal impressive classification precision, instantaneous performance, and practical applicability in academic contexts.

Keywords: Emotion Detection, Deep Learning, ResNet, ZFNet, Real-Time Processing, Student Engagement, Computer Vision

INTRODUCTION

Student motivation is a vital element impacting academic achievement; nevertheless, traditional approaches to measuring engagement depend on manual observation and self-reported questionnaires that are often subjective and inefficient. With advancements in deep learning and computer vision, automated recognition of emotions through facial expressions has surfaced as a promising method for assessing engagement levels in educational environments. Deep learning models possess the capability to evaluate facial features to categorize emotions, generating real-time insights into student motivation levels. However, current models encounter obstacles in attaining high precision and efficiency, particularly in realtime scenarios. The paper at hand introduces a real-time emotion recognition system that utilizes a hybrid deep learning structure, melding ZFNet and ResNet to enhance feature extraction and classification efficiency.

Trained on the EmotiW2019 dataset, the model classifies student emotions based on engagement metrics, such as "engaged" or "not engaged." By scrutinizing facial expressions obtained via a webcam, the system processes the input and provides immediate feedback using visual indicators like emojis and engagement ratings. This flexible and non-invasive method permits ongoing surveillance of student participation in both physical and virtual learning conditions, facilitating tailored interventions to enhance educational experiences.

LITERATURE REVIEW

Facial emotion recognition (FER) has emerged as a highly researched topic within artificial intelligence, with marked advancements fueled by deep learning techniques. Traditional methodologies relied on custom-feature extraction approaches like Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG), which were frequently hindered by their failure to capture intricate facial variations. With the introduction of Convolutional Neural Networks (CNNs), models such as VGG16, MobileNet, and InceptionNet have shown considerable advancements in classifying facial expressions. These architectures utilize profound hierarchical feature extraction yet often struggle with generalization across varied datasets, resulting in inconsistencies in practical applications. He et al. (2016) presented ResNet (Residual Networks), a deep convolutional neural network that addressed the vanishing gradient issue via residual connections, significantly enhancing image classification performance [1]. In a similar vein, Zeiler and Fergus (2014) suggested ZFNet, an improvement upon AlexNet that enhanced convolutional filters and visualization techniques for better feature extraction [2]. Both models have evidenced robust performance in feature learning, establishing them as prime contenders for emotion recognition. Sharma, Dhall, and Goecke (2021) employed deep learning approaches to forecast student motivation levels based on facial expressions, underlining the significance of CNN-based feature extraction [3]. Dhall et al. (2018) introduced the DAiSEE dataset, a benchmark for predicting engagement utilizing facial and ocular features, which has been extensively utilized in affective computing [4]. These studies underline the significance of CNN architectures such as ResNet and ZFNet in the context of emotion classification.

Krizhevsky, Sutskever, and Hinton (2012) created AlexNet, which transformed deep learning in image classification, serving as a catalyst for subsequent architectures like ZFNet and ResNet [5]. Gunes and Piccardi (2009) delved deeper into emotion recognition through facial and bodily expressions, showcasing the potential of multimodal strategies for detecting emotions [6]. These pioneering CNN frameworks have significantly shaped contemporary deep learning-driven emotion recognition systems. Hinton, Vinyals, and Dean (2015) presented knowledge distillation, a method that conveys knowledge from a larger model to a smaller counterpart, allowing for more effective deep learning implementations [7]. Brown et al. (2020) exhibited the strength of transformer-based models in few-shot learning, while Vaswani et al. (2017) introduced the attention mechanism, transforming NLP and computer vision domains [8,9]. Although our research centers on CNN architectures, future developments may investigate attention mechanisms to enhance feature fusion. LeCun, Bengio, and Hinton (2015) delivered an extensive review of advancements in deep learning, highlighting CNNs as the foundation of modern computer vision [10]. These insights bolster our methodology of leveraging ResNet and ZFNet for real-time emotion classification. Moreover, the growing access to datasets like DAiSEE has propelled research in affective computing, fostering innovation in monitoring student engagement.

METHODOLOGY

A. *Existing System*

The current methods for identifying student engagement through deep learning algorithms, such as MobileNet, face numerous challenges. A primary concern is that these approaches necessitate manual feature extraction from images, which can be intricate and influences accuracy. Given that MobileNet is tailored for lightweight applications, it may lack the necessary depth and feature extraction proficiency for engagement detection, resulting in diminished accuracy.

Dis Advantages:

Limited Accuracy – Conventional methods often fail to achieve high accuracy due to restricted feature extraction capabilities.

Elevated Complexity – Manual feature extraction heightens computational complexity and requires further processing**.**

Insufficient Model Depth – MobileNet, being a lightweight framework, is deficient in the depth required to capture nuanced facial expressions essential for detection.

Not Suitable for Larger Datasets – These approaches may struggle to scale effectively when employed with extensive datasets featuring varied facial expressions and engagement levels.

B. *Proposed System*

The suggested system employs Convolutional Neural Networks (CNNs) to tackle the limitations found in the current methods for detecting student engagement. Traditional methods, like MobileNet, frequently contend with manual feature extraction, leading to heightened complexity and decreased accuracy. To surmount these hurdles, the system utilizes a Hybrid Deep Learning Framework that merges ZFNet and ResNet for enhanced feature extraction and classification. This automated feature extraction mechanism negates the necessity for manual involvement, allowing the model to proficiently assess facial expressions and ascertain whether a student is engaged or not.

Advantages:

Quicker Detection – The hybrid model drastically cuts down the time needed for feature extraction and classification.

Enhanced Accuracy – Deep learning models autonomously identify pertinent facial features, resulting in more dependable engagement detection.

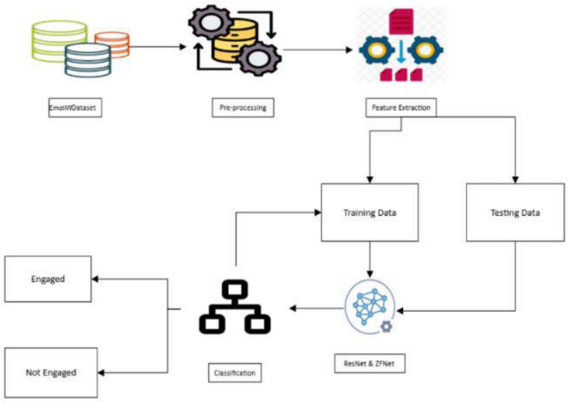

Fig.1.Architecture Diagram of Proposed System

C. *Tools and Platforms used*

The suggested real-time emotion detection system is crafted using a blend of deep learning frameworks and computer vision libraries to guarantee high accuracy and efficiency. The implementation is executed in Python, employing TensorFlow and Keras for deep learning model creation, and OpenCV for image processing operations. Jupyter Notebook acts as the primary development environment. reducing processing time. Additionally, Flask/Dash is used for real-time visualization, allowing for seamless integration of the emotion detection system into a web-based interface for interactive feedback.

D. *Data set used*

Numerous datasets focusing on facial expressions are available. We employed the Facial Expression Recognition dataset (EmotiW2019), which comprises labeled video frames categorized into four levels of engagement: engaged, not engaged,. To enhance model performance, the dataset undergoes a series of preprocessing steps. These steps include resizing images to a standardized dimension for uniformity across training samples, normalizing pixel intensity values for consistency, and applying data augmentation techniques such as flipping, rotation, and brightness modifications to bolster the model's capability to generalize across varying facial expressions and environmental conditions.

Fig.1. Images of the DAiSEE dataset

E. *Data Pre Processing*

In this research, data preparation is crucial for ensuring the optimal training of the deep learning model aimed at emotion recognition. The dataset is initially organized in a hierarchical directory structure, with each subfolder representing a unique emotion class. The quantity of images in each category is tallied to assess class distribution and identify any potential imbalances. To improve model generalization and reduce overfitting, data augmentation practices are implemented, incorporating rotation, width and height shifts, shear transformations, zooming, horizontal flipping, and brightness modifications to introduce variability into the dataset, mimicking real-world scenarios. Additionally, all pixel values are rescaled to the range of 0–1 to standardize inputs and promote efficient learning.To guarantee impartial assessment, the dataset is split into an 80% training portion and a 20% validation portion, with augmentation implemented exclusively on the training data while maintaining the validation portion unchanged. As class imbalance could impede model performance, class weighting is applied by calculating balanced class weights using the compute_class_weight function. These weights are employed during model training to mitigate bias toward overrepresented classes and ensure equitable learning across all categories. This preprocessing strategy markedly enhances the resilience and accuracy of the proposed hybrid deep learning model, significantly improving its capability to detect and categorize emotions with great precision.

F. *Layers and Activation functions used*

1) ***2D Convolutional Layer*** : The two-dimensional convolutional layer, commonly called Conv2D, is an essential component in deep learning, particularly in image processing. This layer extracts features by applying a set of filters to an input image, transforming one 2D feature map into another based on predefined parameters. The effectiveness of this layer depends on the number and size of the filters used. By capturing spatial patterns, Conv2D plays a crucial role in identifying key structures within images.

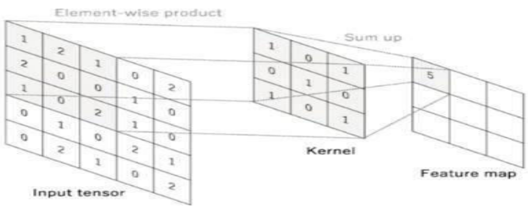

Fig.2. 2D Convolutional operation

Equation :

$$\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sum_{c=0}^{C-1} X(i+m, j+n, c) . W(m, n, c, k) + b_k \quad \text{---eq(1)}$$

Where,

- X(i,j,c) is the input image (feature map) with spatial indices (i,j) and channel c.
- W(m,n,c,k) is the convolution kernel (filter) of size M×N .
- bk is the bias term for filter k.
- Y(i,j,k) is the output feature map at position (i,j) for filter k.
- C is the number of input channels.

2) ***Max Pooling Layer:*** The Max Pooling layer improves feature selection by preserving the single highest value in a particular portion of a feature map. This is done in such a way that the data's dimensionality is reduced while simultaneously keeping the most important features as well as minimizing cost in computation. It also increases efficiency in extracting the most useful information because the input is downsampled which makes Max Pooling an important technique for tasks in computer vision.

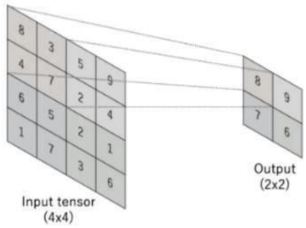

Fig.3.Max Pooling operation

Equation:

$$Y(i, j, c) = \max_{(m,n)\in p} X(P \cdot I + m, P \cdot j + n, c) \text{------} eq\ (2)$$

Where,

- ***P is the pooling window size.***
- ***X(i,j,c) is the input feature map.***
- ***Y(i,j,c) is the output feature map after pooling.***

3) ***Dropout Layer***: The Dropout layer is a form of regularization that helps to avoid overfitting in neural networks. It does this by randomly turning off a percentage of the neurons during training. To ensure that the model does not memorize every data point, it is forced to learn patterns that are more generalized. This enhances the performance and robustness of the network as a whole. Furthermore, generalization can be improved during training by applying early stopping in combination with dropout.

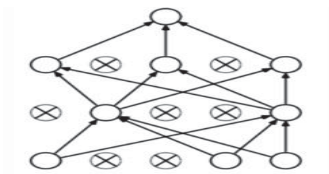

Fig.4. Dropout operation Example

$$Y\,i = \begin{cases} \frac{X\,i}{1-p}, & if\ Xi\ is\ retained \\ 0, & if\ Xi\ is\ dropped \end{cases} \text{---------} eq\ (3)$$

Where,

p is the dropout probability (percentage of neurons dropped).

4) ***Dense Layer**:* A dense layer is synonymously referred to as a fully connected layer in deep learning. The distinguishing characteristic of this layer configuration is that each neuron in it is connected to each neuron in the adjoining layer, thereby allowing matrix-vector multiplication. As deep learning progresses, these layers assist the network in processing and optimizing feature representation matrices. In the majority of neural network architectures, Dense layers enable the combination and analysis of the procured data corresponding to the interpretations.

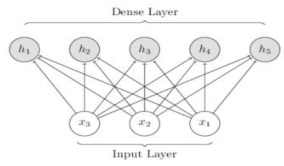

Fig.5.Denseoperation Example

Equation:

$Y = WX + b$ ---------*eq (4)*

Where,

- X is the input vector.
- W is the weight matrix.
- b is the bias vector.
- Y is the output vector

5) ***Flatten Layer**:* The Flatten layer transforms an n-dimensional input tensor into a single dimension. This step guarantees that the pooled feature map will be transformed into a single vector with every pixel mapped to a neuron. In this way, the layer allows the network to consider every feature separately. For example, if the output of a certain layer is a tensor of dimension (4,4,128), the Flatten layer converts it to (2048,1) so that it can be processed by fully connected layers.

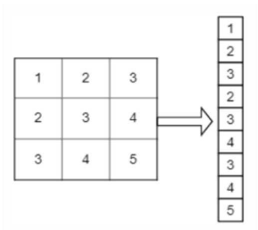

Fig.6.Flatten operation Example

6) ***ReLU Activation***: The Rectified Linear Unit (ReLU) activation function is widely used in deep learning models to introduce non-linearity. It operates by returning zero for negative inputs while allowing positive inputs to pass through unchanged. This mechanism prevents vanishing gradient issues and accelerates training by enhancing the efficiency of weight updates.

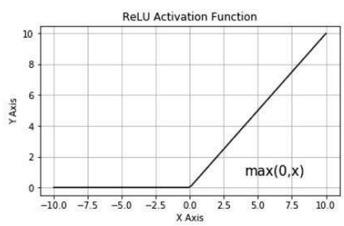

Fig.7.ReLUActivationfunction

Equation:

$$f(x)=\max(0,x) \text{-------- } eq(6)$$

where:

- *If $x \geq 0$, the output is x.*
- *If $x<0$,the output is 0.*

7) ***Softmax Activation**:* The Softmax activation function is primarily used for multi-class classification problems. It transforms the raw network outputs into a probability distribution where the sum of all probabilities equals 1. Inputs with higher values receive larger probabilities, while smaller or negative inputs are assigned lower probabilities, ensuring an interpretable output for classification tasks. Softmax is commonly paired with cross-entropy loss for optimal performance.

Equation:

$$S\ (x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \qquad \text{--------------- } eq\ (7)$$

where:

- Xi is the input value for class iii.
- n is the number of output classes.
- The denominator ensures all outputs sum to 1.

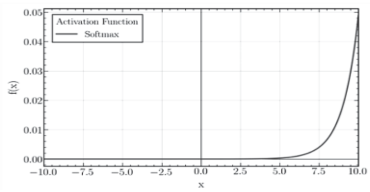

Fig.8.Softmax Activation function

8) ***Adam Optimizer:*** The Adam optimizer is an adaptive optimization algorithm that adjusts learning rates dynamically for each parameter. It integrates features from both Stochastic Gradient Descent (SGD) and RMSprop, utilizing the moving average of gradients for momentum. This approach enhances stability and convergence speed, making Adam a preferred choice for deep learning models.

***9) Cross-Entropy Loss Function**:* Cross-entropy is a crucial loss function used for binary and multi-class classification tasks. It measures the discrepancy between the predicted probability distribution and the actual labels, with lower values indicating better model performance. A perfect model would achieve a cross-entropy loss of zero. This loss function plays a vital role in optimizing neural networks by guiding weight updates toward minimizing classification errors.

ALGORITHM: Hybrid Approach

Input: EmotiW 2019 dataset

Output: Predict the image is engaged or not engaged

Start

{

Initialization

for $\{x_1, x_2, x_3, \ldots, x_n\}$

// Load EmotiW 2019 dataset and preprocess images

Preprocessing

$P \rightarrow X+\gamma \cdot \sigma(V)P$

// Convert images to grayscale, resize, and normalize

FeatureExtractionusingZFNet

$FZF \rightarrow X+WZF[\max(Z,N)]$

// ZFNet extracts hierarchical image features

FeatureExtractionusingResNet

$FRES \rightarrow X+WRES[\max(Z,N)]$

// ResNet captures deep semantic features withresidual connection.

Feature Fusion (Combining ZFNet&ResNet Features)

$FHybrid=\lambda 1FZF+\lambda 2FRES$

// Weighted sum of extracted features for robustness

Feature Optimization using Adam Optimizer

$$w_t = w_{t-1} - \frac{\alpha}{\sqrt{v_t + \epsilon}} M_t$$

// Optimize hybrid features for classification

ExtractedFeatureStoring

$Y \rightarrow D + I\,[X\,\{x_1, x_2, x_3, \ldots\ldots x_n\}]$

// Save hybrid extracted features for training

TestingUnseenImage

$T \rightarrow X-\beta[V]$

// Load new image for classification

Emotion Classification using SoftMax

Compute probability for each emotion class:

$$S(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

MatchingProcess

if$(Q \leq 1)$

{

Emotion detected: Engaged or Not Engaged.
}
else(Q>0)
{
Emotion Not Identified (Request User Input)
}
Return the best solution.
}
End

G. ***Network Training***

The development of our hybrid ResNet-ZFNet model involves extracting features from the pre-trained layers of ResNet50 while incorporating fully connected layers for classification. The model is trained on the EmotiW2019 dataset, where images are subjected to rescaling, augmentation (rotation, flipping, brightness adjustments), and contrast enhancement to enhance generalization. Class imbalance is tackled using calculated class weights, ensuring equitable training across engagement levels. The model is compiled with the Adam optimizer (learning rate = 0.0001) and is trained using categorical cross-entropy loss for multi-class classification. Training is conducted for 25 epochs with batch normalization and dropout layers (0.5) to mitigate overfitting. After training, the model achieves an accuracy of 98%.

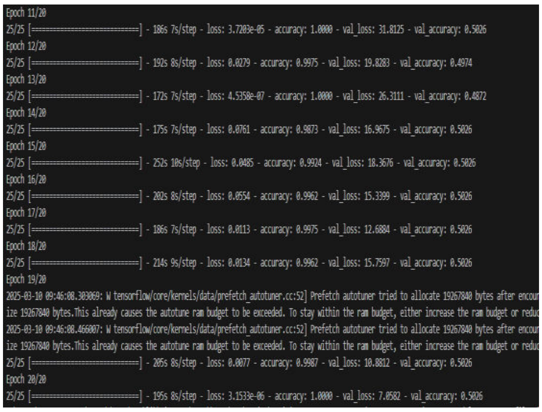

Fig.9.Accuracy of a model

RESULTANDDISCUSSION

REALTIME TESTING

After developing the hybrid ResNet-ZFNet model, it processes live feeds from a webcam, utilizing OpenCV and Dlib for face detection before implementing preprocessing steps such as resizing and normalization. The model categorizes engagement levels into "Engaged" or "Not Engaged" and delivers immediate feedback based on engagement metrics, achieving notable accuracy with minimal latency.

Fig.11. Working of the model

CLASSIFICATION REPORT & CONFUSION MATRIX

The classification report indicates that the model reaches an impressive accuracy of 98%, with excellent precision (0.99 for Engaged, 0.98 for Not Engaged) and recall (0.98 for Engaged, 0.99 for Not Engaged), culminating in an F1-score of 0.98 for both categories. This reflects a well-adjusted model with a low incidence of false predictions. The confusion matrix reveals 95 instances of correctly identified "Engaged" cases, 97 instances of accurately classified "Not Engaged," alongside 2 false positives (Not Engaged mistakenly tagged as Engaged), and merely 1 false negative (Engaged incorrectly judged as Not Engaged). With only 3 misclassifications from 195 predictions, the model showcases outstanding classification capabilities, making it a reliable tool for real-time detection of student emotions.

🔍 Classification Report:

	precision	recall	f1-score	support
Engaged	0.99	0.98	0.98	97
Not engaged	0.98	0.99	0.98	98
accuracy			0.98	195
macro avg	0.98	0.98	0.98	195
weighted avg	0.98	0.98	0.98	195

🟢 Confusion Matrix:

```
[[95  2]
 [ 1 97]]
```

Fig.9.Classification report

The hybrid ResNet-ZFNet model notched an impressive classification accuracy of 98%, highlighting its efficacy in the real-time detection of student emotions. The classification report displays strong measures of precision (0.99 for Engaged, 0.98 for Not Engaged) and recall (0.98 for Engaged, 0.99 for Not Engaged), leading to a harmonized F1-score of 0.98. The confusion matrix confirms just 3 misclassifications throughout 195 predictions, affirming the model's resilience. The real-time testing further substantiated the system's performance, yielding immediate feedback on engagement with negligible delays. Incorporating techniques like data augmentation, batch normalization, and dropout layers significantly reduced overfitting, while class weighting helped tackle dataset disparities. Compared to conventional assessment methods, this deep learning-driven approach presents a scalable, impartial, and automated mechanism for monitoring student attentiveness in both physical and online classrooms.

Conclusion And Future Work

The design phase of the student engagement classification system sets up a sturdy framework for real-time recognition of emotions through facial expressions. Utilizing a hybrid ZFNet-ResNet structure, this system incorporates modular components for tasks such as face detection, preprocessing, feature extraction, and classification of engagement. By correlating emotions with levels of engagement and offering real-time visual feedback via emojis and metrics, the design guarantees scalability and adaptability in both face-to-face and virtual classroom settings. This stage establishes a solid groundwork for subsequent development, striving to enrich educational experiences through actionable insights and tailored interventions. Future endeavors for this real-time facial emotion detection initiative can focus on several pivotal enhancements. A key area for exploration is multi-modal emotion detection, where facial expressions meld with audio-based emotion recognition and physiological signals like heart rate and EEG, facilitating a more comprehensive understanding of emotions. Another focus could be on real-time feedback and adaptation of emotional responses, enabling the system to dynamically modify learning environments based on students' emotional states, thereby providing personalized and adaptive learning experiences. Additionally, the project could find its way into human-computer interaction (HCI) applications, such as virtual classrooms where educators receive real-time insights regarding student engagement or AI-driven chatbots that interact empathetically in response to recognized emotions.

REFERENCES

1. W. Mellouka and W. Handouzia, "Facial emotion recognition using deep learning: Review and insights," in *Proc. 2nd Int. Workshop Future Internet Everything (FIoE)*, Leuven, Belgium, 2020, pp. 1-8.
2. S. Kaur and N. Kulkarni, "Emotion recognition – A review," CSE, MIT SOE, MIT ADT University, Pune, India, 2020.
3. J. A. Ballesteros, G. M. Ramírez, F. Moreira, A. Solano, and C. A. Pelaez, "Facial emotion recognition through artificial intelligence," University of Cauca, Colombia, 2020.
4. S. M. Saleem Abdullah, S. Y. Ameen, M. A. M.sadeeq, and S. R. M. Zeebaree, "Multimodal emotion recognition using deep learning," 2020.
5. S. J. Oh, J. Y. Lee, and D. K. Kim, "The design of CNN architectures for optimal six basic emotion classification using multiple physiological signals," 2020.
6. P. Vanneste, J. Oramas, T. Verelst, T. Tuytelaars, A. Raes, F. Depaepe, and W. Van den Noortgate, "Computer vision and human behaviour, emotion and cognition detection: A use case on student engagement," 2020.
7. H. I. James, J. J. Anto Arnold, J. M. Masilla Ruban, M. Tamilarasan, and R. Saranya, "Emotion-based music recommendation system," Valliammai Engineering College, Chennai, Tamil Nadu, 2020.
8. E. M. Onyema, P. K. Shukla, S. Dalal, M. N. Mathur, M. Zakariah, and B. Tiwari, "Enhancement

of patient facial recognition through deep learning algorithm: ConvNet," 2020.
9. A. Khopkar and A. A. Saxena, "Facial expression recognition using CNN with Keras," DES's Navinchandra Mehta Institute of Technology and Development & Research Scholar Pacific University, Udaipur, India, 2020.
10. A. Agrawal and N. Mittal, "Using CNN for facial expression recognition: A study of the effects of kernel size and number of filters on accuracy," 2020.
11. I.-K. Choi, H.-E. Ahn, and J. Yoo, "Facial expression classification using deep convolutional neural network," 2020.
12. Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), 18-31
13. https://www.researchgate.net/publication/386597306_Real-Time_Emotion_Recognition_for_Improving_the_Teaching-Learning_Process_A_Scoping_Review,2024
14. https://www.ijcaonline.org/archives/volume186/number56/real-time-emotion-recognition-using-deep-learning-a-comprehensive-approach/,2024
15. Khare, S.; Blanes-Vidal, V.; Nadimi, E.; Acharya, U. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. Inf. Fusion 2023, 102, 102019
16. Cai, Y.; Li, X.; Li, J. Emotion Recognition Using Different Sensors, Emotion Models,Methods and Datasets: A Comprehensive Review. Sensors 2023, 23, 2455
17. Taype-Huarca, L.A.; Zavalaga-del Carpio, A.L.; Fernández-González, S.V. Usefulness of the Perezhivanie Construct in Affectivityand Learning: A Systematic Review. In Affectivity and Learning: Bridging the Gap Between Neurosciences, Cultural and CognitivePsychology; Springer Nature: Cham, Switzerland, 2023
18. Cristóvão, A.M.; Valente, S.; Rebelo, H.; Ruivo, A.F. Emotional education for sustainable development: A curriculum analysis ofteacher training in Portugal and Spain. Front. Educ. 2023,8, 1165319
19. Wozniak, M.; Sakowicz, M.; Ledwosinski, K.; Rzepkowski, J.; Czapla, P.; Zaporowski, S. Bimodal Emotion Recognition Based onVocal and Facial Features. Procedia Comput. Sci. 2023,225, 2556–2566.