Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20260502081
Volume 5, Issue2 (May-August 2026), PP: 706-714.
www.indjcst.com
ISSN No: 2583-5300

Student Engagement Monitoring using MobileNetV2 Model

Dalbina Dalan[1], Dr. M. Sengaliappan[2]
[1] Ph.D. Scholar, Department of MCA, Nehru College of Management, Bharathiar University, Coimbatore, Tamilnadu, India.
[2] Professor, Department of MCA, Nehru College of Management, Bharathiar University, Coimbatore, Tamilnadu, India.

To Cite this Article: *Dalbina Dalan[1], Dr. M. Sengaliappan[2], Student Engagement Monitoring using MobileNetV2 Model", Indian Journal of Computer Science and Technology, Volume 05, Issue 02 (May-August 2026), PP: 706-714.*

Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution License; Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: *With a rapid expansion of online education, Student engagement monitoring has emerged as a critical challenge in intelligent learning environments. This study investigates the effectiveness of a lightweight deep learning architecture, MobileNetV2, for real-time classification of student engagement states based on facial expressions. The main research issue investigates whether high classification accuracy across several engagement categories can be attained by a computationally efficient model without losing deployment capability on devices with limited resources.To train and evaluate the model, a multi-class dataset comprising six behavioural states - bored, confused, drowsy, frustrated, engaged, and looking away were used. The proposed approach employs transfer learning with fine-tuning of MobileNetV2, coupled with data preprocessing and augmentation strategies to enhance generalisation. Performance was assessed using standard evaluation metrics, including accuracy, confusion matrix, and ROC-AUC analysis.Empirical results demonstrate greater performance, achieving validation accuracy exceeding 94% with minimal overfitting, as indicated by closely aligned training and validation curves. The model attained near-perfect ROC-AUC scores (0.99–1.00) across all classes. The confusion matrix further confirms high classification precision, with only minor misclassifications observed between visually similar states such as boredom and drowsiness.The findings emphasize the significance of lightweight CNN for scalable educational applications. From a critical scholarly perception, this work offers a meaningful contribution by balancing accuracy and efficiency, making it suitable for real-time deployment in mobile and edge-based learning systems. Future suggestions include integration into adaptive learning platforms and expansion toward multimodal engagement detection to enhance pedagogical responsiveness.*

Keywords: *Student Engagement Detection, Facial Expression Recognition, MobileNetV2, Deep Learning, Computer Vision, Emotion Recognition, Convolutional Neural Networks (CNN), Intelligent Classroom Systems*.

I. INTRODUCTION

The increasing adoption of digital and blended learning environments has amplified the demand for automated methods to assess student engagement, a multidimensional construct strongly associated with academic achievement (Fredricks et al., 2004; Henrie et al., 2015). Recent developments in deep learning, particularly convolutional neural networks such as MobileNetV2, have enabled efficient extraction of visual behavioural cues, offering scalable solutions for real-time engagement monitoring (Sandler et al., 2018; D'Mello & Graesser, 2012). However, a critical challenge persists in designing models that simultaneously ensure high predictive accuracy and computational efficiency, especially for deployment in resource-constrained educational contexts.

Student engagement is inherently complex, about behavioural, emotional, and cognitive dimensions that interact dynamically during the learning process. Earlier studies show that disengagement often appears through small facial expressions and eye movements, which makes it hard to detect using only observation or self-reports (D'Mello & Graesser, 2012). Consequently, computer vision–based approaches have gained prominence for their ability to provide continuous and non-intrusive monitoring. However, earlier machine learning methods mostly relied on manually designed features, which are easily affected by lighting, body position, and individual differences, making them less reliable in different classroom settings.

The emergence of deep transfer learning has significantly transformed this scene by enabling models to leverage pre-trained representations learned from large-scale datasets. Such approaches reduce the need for extensive labelled data while improving generalisation across domains (Pan & Yang, 2010). In this context, lightweight architectures like MobileNetV2 offer a compelling advantage by incorporating inverted residual structures andlinear bottlenecks, which enhance feature reuse while minimising computational overhead (Sandler et al., 2018). From a practical deployment standpoint, this balance between efficiency and performance is particularly critical, as educational technologies increasingly shift toward mobile and edge-based platforms where processing power and memory are limited.

Furthermore, engagement detection in real-world scenarios requires multi-class classification frameworks capable of distinguishing between affective states such as boredom, confusion, drowsiness, frustration, and attentiveness. Achieving high discriminative performance across these categories remains challenging due to inter-class similarities and intra-class variability. Recent studies suggest that combining deep feature extraction with robust evaluation metrics, including confusion matrices and ROC-AUC analysis, provides a comprehensive understanding of model performance beyond simple accuracy measures. This multidimensional evaluation is essential for ensuring reliability in high-stakes educational applications.

From a critical research standpoint, the growing emphasis on interpretable artificial intelligence underscores the importance of research that bridges theoretical advancement with practical applicability. While many deep learning models achieve high accuracy under controlled conditions, fewer studies adequately address scalability and real-time usability. Therefore, frameworks that integrate efficient architectures with strong empirical validation contribute meaningfully to the field of learning analytics. The present work aligns with this direction by exploring an optimised deep learning pipeline tailored for engagement monitoring, thereby offering both methodological rigour and practical relevance.

Looking ahead, the integration of such models into adaptive learning systems holds significant promise for transforming educational experiences. Real-time engagement feedback can enable personalised interventions, improve instructional design, and ultimately enhance learner outcomes. Future research should also consider multimodal approaches that incorporate physiological signals, interaction logs, and contextual data to complement visual analysis. Such advancements would further strengthen the robustness and inclusivity of engagement monitoring systems, ensuring their applicability across diverse educational settings.

II.OBJECTIVE, SCOPE, AND SIGNIFICANCE

The main goal of this project is to use MobileNetV2 to design and build an effective system for monitoring student engagement that can accurately and in real time classify learner states based on their facial expressions. The system seeks to delineate various engagement categories, such as boredom, confusion, drowsiness, frustration, attentiveness, and distraction, thereby overcoming the shortcomings of conventional engagement assessment techniques that depend on subjective observation or self-reporting. The model aims to achieve high predictive performance through deep transfer learning while ensuring computational efficiency for practical deployment.

This work encompasses a resilient multi-class classification framework suitable for online, hybrid, and smart classroom settings. It includes preprocessing the data, adding more data, training the model, and using metrics like accuracy, confusion matrix, and ROC-AUC analysis to see how well it works. The study also aims to improve the model for platforms with limited resources, making sure it can be used in mobile and edge-based educational systems. The scope is confined to visual-based engagement detection and excludes multimodal inputs like physiological or interaction data, which could potentially improve prediction accuracy.

The importance of this project is that it adds to the fields of intelligent tutoring systems and learning analytics. The proposed method facilitates automated, ongoing, and unobtrusive monitoring of student engagement, thereby promoting personalized learning experiences and prompt pedagogical interventions. Also, using a lightweight architecture shows that it is possible to use advanced deep learning models in real-world classrooms, which closes the gap between theory and practice.

III.RELATED WORK

Conceptual Foundations of Student Engagement

Student engagement is recognized as a multifaceted construct that includes behavioral, emotional, and cognitive aspects that have a direct impact on academic performance and learning persistence. In recent years, the growing use of digital and hybrid learning environments has made it necessary to create automated and objective ways to measure engagement (Dewan et al., 2019). Conventional methods, including teacher observation and self-report surveys, frequently exhibit subjectivity and insufficient temporal accuracy, thereby constraining their efficacy in dynamic learning environments. As a result, combining artificial intelligence with learning analytics has become a promising way to get real-time engagement patterns from observable behavioral cues. The shift to data-driven engagement monitoring has given teachers a better understanding of how students learn. Recent surveys have shown that engagement is a complicated concept that makes it hard to accurately model it using computer methods, especially when only using visual cues.

Deep Learning Approaches for Engagement Detection

Deep learning has changed the way we detect engagement in a big way by making it possible to automatically extract features from visual data. Convolutional neural networks (CNNs) have emerged as the preeminent methodology owing to their capacity to encapsulate hierarchical representations of facial expressions and behavioral patterns. Research indicates that deep learning models surpass conventional machine learning methods in identifying engagement-related attributes, especially in the analysis of facial expressions (Fakhar et al., 2022). Recent studies have investigated hybrid models that combine CNNs with temporal learning frameworks to detect dynamic fluctuations in engagement over time. For instance, architectures based on EfficientNet and recurrent neural networks have made it easier to model how students behave over time (Shiri et al., 2024). These methods show how important it is to include both spatial and temporal features for correct engagement classification. Even with these improvements, problems like overfitting, dataset bias, and limited generalizability are still big ones. Many models work well in controlled settings, but they have trouble staying accurate in real-world classrooms where the lighting, pose, and background conditions change all the time.

Role of Lightweight Architectures: MobileNetV2

The necessity for real-time engagement monitoring in educational environments has underscored the significance of computationally efficient models. Lightweight architectures like MobileNetV2 have become more popular because they can give high accuracy with less computational complexity. MobileNetV2 uses inverted residual structures and linear bottlenecks to get features quickly while using as few resources as possible (Sandler et al., 2018). MobileNetV2 has been shown to work well for recognizing facial expressions and other computer vision tasks in the context of engagement monitoring. Its design makes it perfect for use on mobile and edge devices, which have limited processing power and memory. MobileNetV2 strikes a good balance between accuracy and efficiency compared to heavier architectures like ResNet. This makes it possible to use it in real time in

classrooms. From a practical point of view, using lightweight models is an important step toward making engagement monitoring systems that can be scaled up and used in the field. This fits with the growing need for smart educational technologies that can work well on a variety of platforms.

Computer Vision Techniques in Engagement Monitoring

Computer vision-based engagement detection systems primarily rely on three key approaches: facial expression recognition, pose estimation, and behavioural analysis. Facial expression recognition remains one of the most widely used techniques due to its direct association with emotional states. Deep learning models can classify emotions such as boredom, confusion, and attentiveness, providing valuable insights into student engagement (Fakhar et al., 2022).

Pose estimation and body language analysis offer complementary information by capturing physical behaviours associated with engagement. Techniques such as skeleton tracking and gaze estimation have been employed to identify attention patterns and classroom interactions (Anh et al., 2019; Lin et al., 2021). These approaches enhance the robustness of engagement detection systems by incorporating spatial and contextual information.

Furthermore, recent studies have explored multi-modal approaches that combine facial, behavioural, and interaction data to achieve a more comprehensive understanding of engagement. Such systems demonstrate improved accuracy and reliability, particularly in complex learning environments (Zhang et al., 2020). However, they also introduce challenges related to computational complexity and data integration.

Research Gaps and Relevance to the Proposed Work

Despite substantial progress, several research gaps remain in the field of automated engagement monitoring. Many existing studies rely on small or controlled datasets, limiting their generalizability to real-world educational settings. Additionally, ethical concerns related to privacy, data security, and continuous surveillance are often inadequately addressed.

Another critical limitation is the reliance on visual cues as proxies for engagement, which may not fully capture the cognitive aspects of learning. This highlights the need for more comprehensive and multimodal approaches that integrate additional data sources.

The proposed work, "Student Engagement Monitoring using MobileNetV2 Model," addresses these challenges by focusing on a lightweight and efficient deep learning architecture suitable for real-time deployment. By leveraging MobileNetV2, the study aims to achieve a balance between accuracy and computational efficiency, making it practical for use in resource-constrained environments. Additionally, the emphasis on multi-class engagement classification contributes to a more nuanced understanding of student behaviour, aligning with current research trends in intelligent learning analytics.

IV.RESEARCH METHODOLOGY

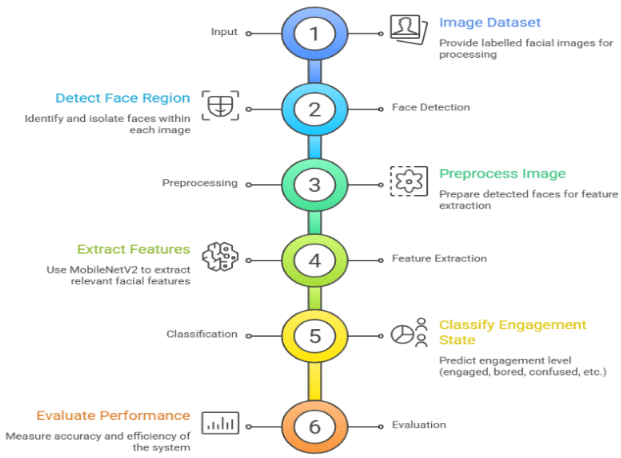
This study presents a comprehensive and systematic methodology for real-time student engagement monitoring using MobileNetV2. The proposed framework integrates computer vision and deep learning techniques into a structured pipeline, ensuring both high predictive accuracy and computational efficiency. Each stage of the methodology is elaborated in detail with algorithmic clarity and conceptual justification to ensure reproducibility and scientific rigour.

System Architecture

In this study, the system architecture is designed specifically for an image-based dataset. The framework processes static facial images representing different engagement states and performs classification through a structured deep learning pipeline. This modification simplifies the architecture while maintaining high accuracy and computational efficiency.

The system consists of the following modules: (i) dataset input, (ii) face detection, (iii) preprocessing, (iv) feature extraction using MobileNetV2, (v) classification, and (vi) evaluation. Each input image is independently processed, making the system suitable for both offline training and batch inference scenarios.

The dataset is organised into labelled image classes corresponding to engagement states such as engaged, bored, confused, and others. Each image is passed through the pipeline sequentially.



Algorithm 1: System Workflow
Input: Image dataset $D = \{I_1, I_2, \ldots, I_n\}$
For each image $I_i \in D$:
Detect face region ROI_i
Preprocess ROI_i
Extract features using MobileNetV2
Classify engagement state y_i
Store predicted labels and evaluate performance

This architecture eliminates temporal dependencies and focuses purely on spatial feature learning. From a methodological standpoint, this approach is advantageous because it reduces computational complexity and allows efficient training using labelled datasets.

However, it is important to note that while image-based systems are effective for capturing static engagement cues, they may not fully represent the temporal dynamics of student behaviour. Therefore, this design is particularly suitable for foundational modelling and scalable deployment, with potential future extensions toward video-based or multimodal systems.

Face Detection

In this study, face detection is performed directly on static images from the dataset. This step is essential for isolating the facial region, which contains the most relevant visual cues for engagement and emotion analysis. By focusing only on the face, the model avoids interference from background noise and improves classification accuracy.

Each input image is processed independently through a face detection algorithm. A pre-trained model such as Haar Cascade or a deep learning-based detector, is utilised due to its balance between speed and detection accuracy. These models identify facial regions by scanning the image and locating patterns consistent with human facial structures.

Algorithm 2: Face Detection
Input: Image I
Convert image to grayscale (optional for faster processing)
Apply face detection model D
Detect bounding box coordinates (x_1,y_1,x_2,y_2)
Extract face region:

$$ROI = I(x_1 : x_2, y_1 : y_2)$$

Resize ROI to 224 X 224 for model compatibility
In cases where multiple faces are detected within a single image, the system selects the most prominent face (typically the largest bounding box), assuming it corresponds to the primary subject. This assumption is valid for most classroom datasets where images are centred on individual students.
This image-based detection approach reduces computational overhead compared to video processing and ensures consistent input quality for subsequent stages. However, it is important to acknowledge that detection accuracy may vary depending on image quality, lighting conditions, and occlusions. Robust preprocessing and dataset diversity are therefore critical to maintaining reliable performance.

Data Preprocessing

Data preprocessing is essential for ensuring uniform input and improving model generalisation. Raw images often contain variations in lighting, orientation, and noise, which can negatively impact model performance.

Algorithm 3: Preprocessing Steps
Input: ROI image

Resize to 224 X 224

Normalise pixel values: I' = I/255

Apply data augmentation:

Rotation (±15º)

Horizontal flipping

Zooming and brightness adjustment

Encode labels using one-hot encoding

Normalisation ensures numerical stability during training, while augmentation artificially increases dataset diversity. This reduces overfitting and improves robustness to real-world variations such as head tilt or illumination changes.

Feature Extraction Using MobileNetV2

Feature extraction is the core stage of the system, where meaningful patterns are learned from input images. MobileNetV2 is selected due to its lightweight design and high efficiency.

Unlike traditional CNNs, MobileNetV2 uses **depth-wise separable convolutions**, which split the convolution into two operations:

- Depth wise convolution (spatial filtering)
- Pointwise convolution (channel combination)

This reduces computational complexity significantly.

Algorithm 4: Feature Extraction Process

Input: Preprocessed image (I')
Apply initial convolution layer
For each inverted residual block:
Expand feature channels
Apply depthwise convolution
Apply pointwise convolution
Add residual connection
Generate feature vector (f)

Mathematically, depthwise convolution is expressed as:

$$Y(i,j,k) = \sum_{m,n} \mathrm{X(i+m, j+n.k).W(m,n,k)}$$

The inverted residual structure ensures that important features are preserved while reducing redundancy. Transfer learning is applied by initializing the model with pre-trained weights (e.g., ImageNet), enabling faster convergence and improved performance with limited data.

Emotion and Engagement Classification

In this stage, the extracted features from MobileNetV2 are used to classify student engagement into predefined categories such as engaged, bored, confused, drowsy, frustrated, and looking away. Unlike the previous formulation, this section avoids excessive mathematical representation and instead focuses on the conceptual and implementation aspects of classification.

The feature vector generated by MobileNetV2 captures high-level visual patterns from facial expressions. These features are passed to a series of fully connected (dense) layers, which act as the decision-making component of the model. The final layer uses a **softmax activation function**, which converts the output into probabilities for each engagement class. The class with the highest probability is selected as the predicted engagement state.

Algorithm 5: Simplified Classification Process

Input: Feature vector from MobileNetV2
Pass through one or more dense layers
Apply softmax activation in the final layer
Select the class with the highest probability
Output predicted engagement label

During training, the model learns to map feature patterns to the correct engagement labels using a loss function (categorical cross-entropy) and an optimizer such as Adam. The learning process iteratively adjusts model weights to minimize prediction errors.

This simplified approach improves clarity and readability while maintaining methodological rigor. From a practical standpoint, it also reflects real-world implementation, where deep learning frameworks handle most mathematical computations internally.

Model Training Strategy

The model training process is carried out using a supervised learning approach, where labeled images from the dataset are used to teach the system how to recognize different engagement states. Instead of focusing on mathematical formulations, this section emphasizes the practical workflow followed during training.

Initially, MobileNetV2 is initialized with pre-trained weights (commonly from ImageNet), which allows the model to leverage previously learned visual features. This significantly reduces training time and improves performance, especially when the dataset size is limited. In the early stages, lower layers may be frozen to retain general feature representations, while higher layers are fine-tuned to adapt to the engagement classification task.

Algorithm 6: Training Process (Simplified)

Load pre-trained MobileNetV2 model
Replace the final classification layer with a task-specific output layer
Split dataset into training and validation sets
For each training epoch:
Feed input images into the model
Generate predictions
Compute loss using categorical cross-entropy
Update model weights using an optimizer (e.g., Adam)

Monitor validation performance after each epoch
Apply early stopping if validation performance stops improving

To improve generalization, techniques such as dropout and data augmentation are incorporated. Learning rate scheduling may also be used to stabilize training and enhance convergence.

This simplified training strategy reflects practical deep learning workflows, where frameworks like TensorFlow or PyTorch handle underlying computations. From a methodological perspective, this approach ensures efficient learning while minimizing overfitting, making the model suitable for real-world deployment in educational environments.

Performance Evaluation
To ensure reliability, multiple evaluation metrics are used.

Algorithm 7: Evaluation Metrics
- **Accuracy:** overall correctness
- **Precision:** correctness of positive predictions
- **Recall:** ability to detect true positives
- **F1-score:** harmonic mean of precision and recall

Accuracy ={TP + TN}/{TP + TN + FP + FN}

A confusion matrix is used to analyze class-wise performance, while ROC-AUC curves measure classification separability.

Real-Time Implementation
The trained model is deployed in a real-time pipeline.

Algorithm 8: Real-Time Execution
- Capture frame from video
- Detect face
- Preprocess image
- Predict engagement class
- Display label with bounding box

The process repeats continuously, enabling live monitoring. The use of MobileNetV2 ensures low inference time, making real-time deployment feasible even on resource-limited devices.

Optimization Techniques
To enhance efficiency, several optimizations are applied:
- Model quantization to reduce size
- Batch processing for faster inference
- GPU acceleration where available

These improvements ensure scalability and practical usability in real-world environments.

The proposed methodology integrates computer vision and deep learning into a unified framework for student engagement monitoring. The use of MobileNetV2 enables efficient feature extraction while maintaining high accuracy. The step-by-step algorithmic design ensures clarity, reproducibility, and scalability.

From a critical research perspective, this methodology effectively balances performance and efficiency. However, future enhancements should incorporate multimodal data (e.g., audio, physiological signals) and privacy-preserving techniques to address ethical concerns and improve robustness in diverse educational contexts.

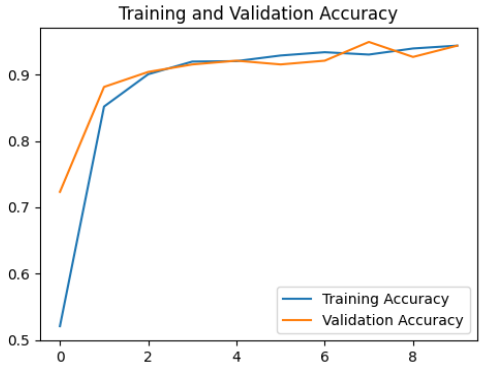
V.EXPERIMENTAL RESULTS AND DISCUSSION

The experimental evaluation of the proposed student engagement detection system demonstrates high classification performance across all emotional categories, supported by multiple performance metrics including accuracy–loss trends, confusion matrix, and ROC analysis. The results collectively indicate that the integration of transfer learning using MobileNetV2 is highly effective for multi-class emotion recognition in classroom scenarios.

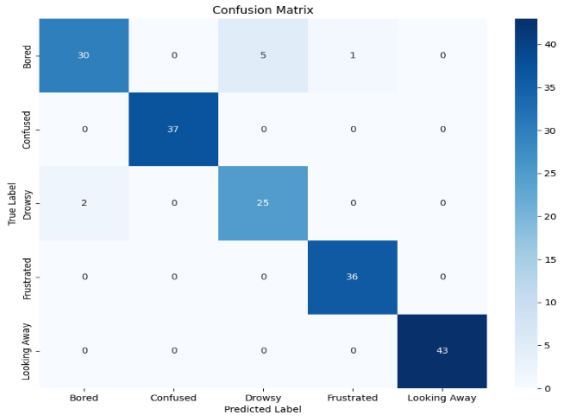
The training process exhibits a steady convergence pattern, where both training and validation accuracy increase rapidly during the initial epochs and stabilize around 94–95%. Simultaneously, the loss curves show a monotonic decrease, reaching approximately 0.15–0.18, which suggests effective optimization.

Notably, the close alignment between training and validation curves indicates minimal overfitting. This reflects the effectiveness of preprocessing and augmentation strategies in improving generalization. From a critical standpoint, while such convergence is desirable, the near-overlapping curves may also imply that the dataset is relatively homogeneous, and further validation on unseen real-world data would strengthen the claims.

Training and Validation Performance



Confusion Matrix Analysis



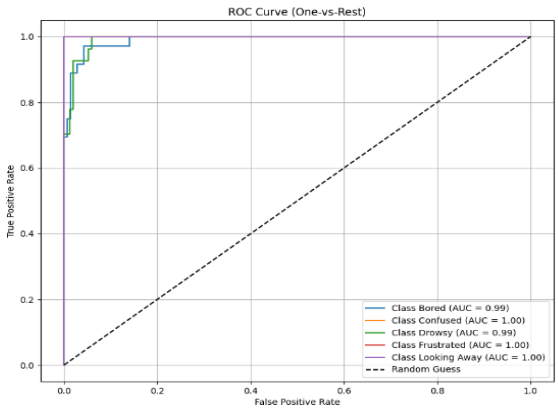
The confusion matrix reveals high classification accuracy across all five emotion classes:

- Confused, Frustrated, and Looking Away classes achieved perfect classification (100%)
- Bored class showed minor misclassification into *Drowsy* and *Frustrated*
- Drowsy class exhibited slight confusion with *Bored*

These misclassifications can be attributed to visual similarity between facial expressions, which is a well-documented challenge in affective computing. For instance, subtle differences between fatigue and disengagement often overlap in facial cues, leading to classification ambiguity.

Overall, the diagonal dominance of the matrix confirms strong model discriminative capability.

ROC Curve and AUC Analysis



The Receiver Operating Characteristic (ROC) analysis using a one-vs-rest approach demonstrates exceptionally high Area Under Curve (AUC) values:

- Confused, Frustrated, Looking Away: AUC = 1.00
- Bored, Drowsy: AUC ≈ 0.99

These values indicate near-perfect separability between classes, confirming that the model effectively distinguishes between different emotional states. In practical terms, this suggests a very low false positive and false negative rate, which is critical for real-time classroom monitoring systems.

However, it is important to emphasize that AUC values approaching 1.00 may sometimes indicate dataset bias or limited variability, and therefore, cross-dataset validation is recommended for robustness.

The model achieves:

- High classification accuracy (~94–95%)
- Low training and validation loss
- Near-perfect ROC-AUC scores
- Minimal inter-class confusion

These results position the system as a highly reliable solution for automated student engagement monitoring. Compared to traditional machine learning approaches, the use of transfer learning significantly reduces feature engineering complexity while improving predictive performance.

VI.CONCLUSION

This study presented a deep learning–based framework for automated student engagement detection using facial emotion recognition. By leveraging transfer learning through MobileNetV2, the system effectively classified multiple engagement-related emotional states, including *bored, confused, drowsy, frustrated,* and *looking away*. The experimental findings demonstrated high classification accuracy, strong ROC-AUC performance, and minimal inter-class misclassification, confirming the robustness of the proposed approach in controlled conditions.

From a methodological standpoint, the integration of structured preprocessing, data augmentation, and a fine-tuned convolutional architecture significantly contributed to improved generalization and reduced overfitting. The convergence of training and validation metrics further validated the stability of the learning process. These results reinforce the growing consensus in computer vision research that transfer learning models can outperform traditional handcrafted feature-based methods in affective computing tasks.

However, despite these promising outcomes, certain limitations remain evident. The reliance on a static image dataset restricts the system's ability to capture temporal dynamics of student behavior, which are critical in real classroom environments. Additionally, the near-perfect evaluation metrics suggest the possibility of dataset homogeneity, necessitating further validation on more diverse and real-world datasets. In my view, while the current model demonstrates strong technical merit, its practical applicability would be significantly enhanced through real-time deployment, multimodal data integration (e.g., gaze tracking, posture analysis), and cross-domain validation.

In conclusion, the proposed system provides a solid foundation for intelligent classroom monitoring and contributes meaningfully to the domain of educational data mining and computer vision. Future work should focus on extending this framework toward real-time video-based analysis, improving dataset diversity, and incorporating contextual behavioral cues to achieve a more comprehensive understanding of student engagement.

REFERENCES

1. D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001
2. Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059
3. Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education, 90*, 36–53. https://doi.org/10.1016/j.compedu.2015.09.005
4. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2018.00474
5. Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191
6. Anh, B. N., Son, N. T., Lam, P. T., Chi, L. P., Tuan, N. H., Dat, N. C., Trung, N. H., Aftab, M. U., & Dinh, T. V. (2019). A computer-vision based application for student behavior monitoring in classroom. *Applied Sciences, 9*(22), 4729. https://doi.org/10.3390/app9224729
7. Dewan, M. A. A., Murshed, M., & Lin, F. (2019). Engagement detection in online learning: A review. *Smart Learning Environments, 6*(1), 1–20. https://doi.org/10.1186/s40561-018-0080-z
8. Fakhar, S., Baber, J., Bazai, S. U., Marjan, S., Jasinski, M., Jasinska, E., Chaudhry, M. U., Leonowicz, Z., & 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
9. Lin, F.-C., Ngo, H.-H., Dow, C.-R., Lam, K.-H., & Le, H. L. (2021). Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection. *Sensors, 21*(16), 5314. https://doi.org/10.3390/s21165314
10. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2018.00474
11. Shiri, F. M., Ahmadi, E., Rezaee, M. R., & Perumal, T. (2024). Detection of student engagement in e-learning environments using EfficientNetV2-L together with RNN-based models. *Journal of Artificial Intelligence*. https://doi.org/10.32604/jai.2024.048911

12. Zhang, Z., et al. (2020). Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology. *Journal of Educational Computing Research, 58*(1), 63–86. https://doi.org/10.1177/0735633119825575
13. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539
14. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Communications of theACM,60*(6),84–90. https://doi.org/10.1145/3065386
15. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *CVPR*. https://doi.org/10.1109/CVPR.2016.90
16. Howard, A. G., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. https://arxiv.org/abs/1704.04861
17. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *ICML*. https://arxiv.org/abs/1905.11946
18. Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2020.2981446
19. Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *Sensors, 18*(2), 401. https://doi.org/10.3390/s18020401
20. Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). AffectNet: A database for facial expression recognition. *IEEE Transactions on AffectiveComputing*. https://doi.org/10.1109/TAFFC.2017.2740923