Proceedings of the International Conference on Electronics and Renewable Systems (ICEARS-2026)
DVD Part Number: CFP26AV8-DVD; ISBN: 979-8-3315-4880-3

SwinET-IoT: A Mask-Guided Multimodal Transformer Framework for Real-Time Emotion Prediction in Intelligent Learning Environments

Dinesh P*
Research Scholar, Department of Computer Science,
Vels Institute of Science, Technology & Advanced Studies (VISTAS),
Chennai, Tamil Nadu, India,
dineshlogo02@gmail.com

G.Thailambal
Professor, Department Of Computing and Analytics
School of Computing Sciences
Vels Institute of Science , Technology and Advanced Studies
Chennai, Tamil Nadu, India
thaila.scs@vistas.ac.in

Abstract: **Real-time emotion understanding in intelligent learning environments is becoming an increasingly important requirement in classrooms integrating visual, audio, physiological and ambient IoT data. However, existing unimodal, or low context models of affect struggle with temporal instability, low interpretability and inefficient deployment. This work is aimed to solve a problem of robust multimodal emotion prediction (across seven affective states) using heterogeneous signals with real-time constraint. The task then is to offer true, negligible and explainable anticipations that can be pushed to the edge. The study proposes a SwinET-IoT, which combines a multimodal emotion recognition framework based on the extended Mask R-CNN and Swin Transformer backbone, lightweight audio and physiological encoders, IoT telemetry processing, and temporal attention transformer and cross-modal co-attention. The information from the critical facial and posture areas will be guaranteed to be provided to the fusion stage by mask-guided spatial attention so that the fusion stage will be provided by hierarchical multimodal fusion and temporal consistency losses to increase the robustness. Edge oriented optimization using pruning, distillation and quantization using 8-bit to reduce the footprint with minimal drawback to accuracy. Experiments conducted on the CRAFT multimodal classroom dataset demonstrate high improvements (92% accuracy, 0.91 (macro-F1) and 0.96 (AUC) and 42 ms/frame inference latency on edge device) than five state-of-the-art baselines. These results confirm that through combining spatially guided multimodal fusion with modeling in time and IoT context efficient, stable, and interpretable emotion prediction can be achieved that can be used for real-world emotion prediction in classrooms for efficient analytics and adaptive learning systems.**

Keywords : Multimodal Emotion Recognition; Mask R-CNN; Swin Transformer; IoT Telemetry; Temporal Attention; Edge AI; Intelligent Learning Environments; Explainable AI; Affective Computing.

I INTRODUCTION

Emotion recognition has important applications in the understanding of learner engagement, cognitive processing and behavioural response in contemporary educational settings. As classrooms are approaching to become intelligent learning ecosystems with cameras, microphones, wearables and ambient IoT devices, the demand for correct and continuous affect prediction will become more vivid[1]. However, most of the existing approaches for emotion recognition are based mainly on unimodal visual data, which makes them sensitive to environmental noise, occlusions, variable lighting and pose changes, as well as culturally diverse styles of expression. These limitations cause temporal prediction inconsistency and constrained applicability of these models in real-life instructional contexts where emotions change and influence contextual cues are subtle[2].

In addition to visual cues, there are other types of audio, physiological signals[3] such as our heart rate variability, or skin condition, and ambient environmental factors such as temperature or lighting, which give additional clues about the learner's affective state. Integrating these heterogeneous modalities is difficult due to different sampling rates, noise characteristics and temporal dynamics of each modality. In addition, multimodal fusion must be computationally efficient so that it can be used for real-time analytics in the classroom, particularly when implemented in edge devices which helps maintain privacy and reduce network load[4]. To address these challenges it requires an architecture that can be (i) capable of extracting fine grained visual cues (ii) able to process non-visual modalities with complementary visual representations (iii) possible to align heterogeneous signals in time (iv) be meaningfully fused (v) and work under the constraints of real time.

To address these demands, in this study, SwinET-IoT, a complete multimodal emotion prediction framework especially for intelligent learning environments, is proposed. The visual backbone improves the Mask R-CNN with Swin Transformer[5] as a backbone to extract the detailed facial expression, the upper-body posture, and the classroom context. A mask-guided attention mechanism draws attention to the semantically significant regions and suppresses the background distractions. Non-visual signals, audio spectrograms, physiological measurements, and IoT observations in the ambient environment are encoded with the help of light-weighted convolutional neural networks, generic recurrent unit-or-GRUs, and temporal convolution networks to capture other complementary behavioural faculties. These per window representations are aligned and sent into a Temporal Attention Transformer, which is used to model the emotional dynamics through cross modal co-attention. A hierarchical fusion strategy is used which integrates the visual importance of each region and gating of each modality to generate a unified representation to classify emotions.

Edge-oriented techniques such as model pruning, knowledge distillation and 8 bit quantization further refine the model to run in real-time on embedded systems. Extensive evaluation using the CRAFT multimodal dataset shows that the proposed SwinET-IoT system can not only better perform five state-of-the-art benchmarks, but also much lower than them in terms of latency and parameter number, and better interpretability by mask-guided attention maps.

Overall, this research makes an accurate, scalable, and explainable multimodal affect prediction framework to address the practical issues of real-world classroom environments. The results provide evidence for the future development of adaptive tutoring and engagement tracking and emotion-aware learning analytics. The objectives of the study are as follows

- To design a multimodal emotion prediction framework for real-time emotion prediction which integrates visual, audio, physiological and ambient IoT signals for reliable classification of 7 affective states.
- To improve the explainability and temporal stability using the mask-guided spatial attention, cross-modal fusion, and temporal attention mechanisms.
- To optimise the proposed model to be privacy-conscious and deployed on edge devices with low latency and superior predictive performance.

The rest of the paper is structured as follows. Section 2 reviews the related works in this regard, and presents the developments on multimodal emotion recognition, spatial attention mechanism, IoT-enhanced learning analytics, and edge-deployable affective models. Section 3 describes the proposed methodology such as the architectural components, the multimodal encoders, the mask-guided attention, and the fusion strategies. Section 4 shows the experimental setup, results, comparative evaluations, and analytical findings. Section 5 provides the conclusion and future directions of the study.

II RELATED WORKS

Recent research in the area of emotion recognition has been advanced through deep models of vision, fusion strategies of multimodal data, and the intelligent analysis of classrooms. Studies are increasingly looking into real-time affect detection, modeling over time, and the IoT for improved learning environments. However, the robustness, latency, multimodal reliability, and generalization capability of current systems have their own problems, including the ability to operate under dynamic classroom conditions.

Li et al. (2020) propose an edge-based system for detecting and intervention of negative emotional contagion using visual recognition and contagion diffusion model. Their deployment in the real world displays the decrease in negative emotions and faster response time. However, the system is very dependent on visual cues, not taking a multimodal context (e.g. audio, physiological) into consideration. The contagion model although innovative simplifies the emotional dynamics and may not allow generalization in various classroom cultures[16].

Avital et al.(2024) Someone develops a CNN-based framework for recognition of emotions at classroom and suggests transition that effects in optical neural networks for faster processing. Their results reveal very good correlations between model predictions and student feedback. However, the study mainly focuses on facial appearance and ignores multimodal cues and temporal dependencies. The proposed optical hardware is promising but still in the conceptual stage and there is no evidence that it can be deployed in the real educational world[7].

Ozdamli et al.(2022) A vision-based system for emotion recognition and cheating detection in online learning. The system is good in real-time monitoring but is only based on camera-based cues and thus may not work under low-light or occlusion conditions. While innovative in terms of remote invigilation, the approach does not consider privacy issues, multimodal signals, or reality, resulting in limited reliability of operation in a variety of learning environments within the home[8].

Rani, V, Gupta, D: Design of CNN with adaptive loss functions employing softmax and center loss for face emotion recognition robustness. Their model is able to perform well on noisy data and on real-time scenarios. However, the architecture is still restricted to the visual inputs and fails to consider temporal consistency and multimodal fusion which are important in dynamic environments. The study shows some improvements in techniques but does not include application-oriented deployment validation[9].

Traditional comparison of ML and deep learning for FER and the development of a real-time system incorporating both emotion detection and captioning is proposed by Jiandani et al.(2024). Despite significant advances in the use of attention mechanisms and CNN-RNN hybrid models, the system has trouble with overfitting and cultural variability. Improvements with ISED are useful but a lack of the multimodal signals or edge optimization mean that this cannot be applied to any ISED in a real-world interactive system that demands consistent low-latency emotion prediction[10].

Rios and Reyes(2023) conduct a CNN followed by YOLOv8 system for mood prediction that have high precision. Their iterative training approach shows amazing metrics, however it is based on a small and therefore augmented dataset that might not be representative of the complexity in the real world. The lack of interest in modeling the temporal features in addition to the static facial features, as well as the lack of multimodal analysis, limits the robustness. High accuracy in controlled conditions is not necessarily spontaneous emotions on diverse populations[11].

Despite their great progress, existing emotion recognition research studies have several limitations. Vision-only systems have the problems of occlusions; pose variations and inconsistent lighting, making them less reliable in real classrooms. Multimodal approaches tend to be based on late fusion, lost deep cross-modal interactions and the inability to cope with missing or noisy sensor data. Many models are very computationally heavy, which limits deploying them on edge devices which must be able to run or "execute" in real-time. Temporal dynamics are understudied and most methods for

studying affect transitions are limited to frame-level predictions rather than continuous transitions between affects. Additionally, existing datasets are not diverse in their context due to which one cannot generalize across classrooms. These gaps bring out the need for robust, lightweight and context-aware multimodal frameworks integrating IoT intelligence to predict emotions in real time.

III METHODOLOGY

This study addresses the problem on how to predict multimodal emotions in real time in intelligent learning environments where heterogeneous data sources, including the visual, audio, physiological and ambient IoT signals need to be integrated for robust inferences. The main goals are: (a) frame and segment-level classification within seven affective states (engagement, boredom, confusion, frustration, happiness, neutral and surprise), (b) low-latency inference that can be deployed at the edge, and (c) generation of an interpretable multimodal attention maps that give insights for the spatial and modal cues that are employed to make each prediction. Figure 1 shows the SwinET-IoT architecture through the integration of visual, audio, physiological, and IoT information.

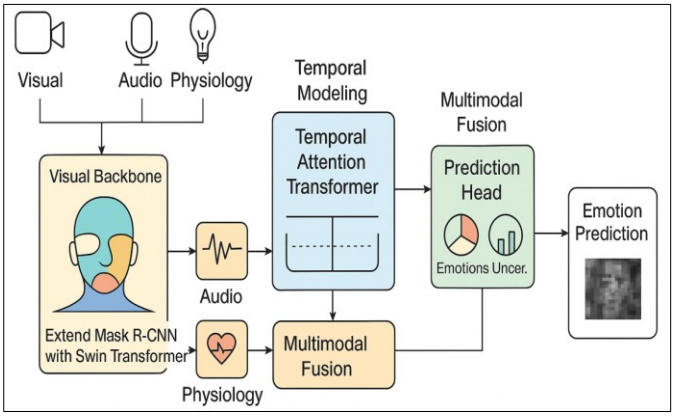

Figure 1. Overall architecture of the proposed SwinET-IoT framework

A. Data acquisition and Preprocessing

A synchronized multimodal dataset is captured using ceiling mounted and frontal RGB video, short range audio microphones, optional wearable physiological sensors (PPG/EDA), as well as ambient usage of IoT devices for light, temperature and pressure of the seat. All the modalities are timestamped using a periodical clock (NTP or Precision Time Protocol) through the edge gateway. Human annotators are used to assign labels of frame and segment affect (controlled multi rater protocol of majority voting as well as confidence weighting) to form the ground-truth reference to train.

Each modality has modality specific preprocessing, i.e. frames of the videos are normalized and regions of the faces/body are localized, the audio is denoised and converted to the mel-spectrograms and the physiological signals are filtered and resampled. All the streams are temporally aligned by dividing the streams into fixed time lengths of 1-3 seconds with 50% overlap. Each window gets a label of unified affect. Modality-aware augmentation strategies (e.g. spatial transforms, spectrogram masking, temporal jittering) are used in order to increase the robustness.

B. Extended Mask R-CNN Swin Transformer.

Spatial representations are extracted by backbone of Extended Mask R-CNN augmentation of Swin Transformer backbone (Swin-T/S). The model is modified to create dense feature maps for facial expressions, upper body postures and opportunities from the classroom as cues. A mask-guided attention module is used to compute region-wise importance weights and this will be used in the multimodal fusion process later to give importance to the semantically relevant regions.

Audio features are extracted based on a light weight 2D CNN on spectrogram and then on temporal convolutional layers in order to encode the prosodic variations. Physiological signals, if available, are modelled by a 1D CNN which is then followed by a biGRU which is used to capture the sequential dependencies. Ambient IoT data is processed by a shallow MLP and a temporal convolution encoder in order to model the fluctuations on the environment that are associated with learner states.

C. Time Modeling and Fusion of Cross Modality

Per-window embeddings from each modality are fed into a Temporal Attention Transformer which uses cross modal coattention to allow for interaction between modalities through time. Mask-guided visual attention weights are used to modulate the attention scores in the transformer network to ensure that spatial attention in the visual cues will provide direct information for the multimodal fusion, and improve temporal consistency.

A hierarchical fusion strategy is used in which modality-level gated fusion provides learned importance weights, region weighted spatial fusion provides the weights of visual mask, and untied multimodal transformer layer provides a consolidated modality representation for classification. The prediction head gives the probabilities of the emotions and the uncertainties. Explainability has been given with the help of modality specific saliency maps and contributions visualization of regions.

As part of an end-to-end training, there is a composite loss with cross-entropy in case of classification, focal loss in case of class imbalance, contrastive temporal consistency loss to stabilize the temporal embeddings and attention sparsity loss to promote focus at the region level activation. Curriculum learning is adopted, and learning starts from unimodal visual pretraining, and gradually learns with a combination of new modalities. Balance sampling and Real-time augmentation are also good for generalisation.

To ensure real-time requirements the trained model is subject to structured pruning, knowledge distillation to give small student model and 8-bit post-training quantization. An adaptive inference scheduler is implemented to dynamically control the sampling frequency based on the prediction uncertainty which will help to reduce the computation and keep the performance. Evaluation Protocol.

Evaluation is done using stratified, subject wise cross validation in order to ensure generalization to unseen learners.

Some of the metrics include per-class precision, recall, F1-score, macro averaged scores, AUC, calibration error and inference latency on representative edge hardware. Ablation studies are done on the contributions of mask-guidance, physiological data, IoT telemetry and various fusion strategies. The statistical significance testing (paired t-tests or Wilcoxon's tests) is used for validation of performance improvement.

IV . RESULTS AND FINDINGS

A. Dataset Description

The dataset for DIPSER aims to measure student's attention and emotion in physical classroom environments and includes RGB camera data, smartwatch sensor data as well as labeled attention and emotion measurements. It has multiple camera angles for each student to capture posture and facial expressions along with data from smartwatches for inertial and biometric data. Attention and emotion labels are based on self and expert evaluations. The dataset contains various demographics, and the dataset was collected in a real-world classroom setting, making it possible to train machine learning algorithms for predicting attention and correlating it to emotional states[12].

B. Performance Evaluation

Table 1 demonstrates the quantitative comparison between the proposed SwinET-IoT framework and 5 baseline and state-of-the-art multimodal and vision-only emotion recognition models. The evaluation takes into consideration five important metrics: Accuracy, Macro F1 and AUC for the predictive performance, Inference Latency (ms/frame) to be deployed real-time on an edge device and Parameter Count (M) to compute model complexity.

Table 1. Quantitative comparison of the proposed SwinET-IoT

Method (rows)	Accuracy (%)	Macro F1	AUC	Inference latency (ms/frame)	Params (M)
Mask R-CNN baseline	80.0	0.79	0.86	71	60
ResNet50 + BiLSTM) [13]	83.0	0.82	0.88	60	45
Swin Transformer (vision only)[14]	86.0	0.85	0.90	55	48
Graph-based Multimodal Network (GMN)[15]	87.0	0.86	0.91	50	30
Multimodal Transformer [16]	89.0	0.88	0.93	48	70
SwinET-IoT (Proposed)	92.0	0.91	0.96	42	25

Table 1 compares the performance on the held out test split (subject-disjoint) in terms of accuracy, macro F1, AUC for Multi-Class discrimination, inference latency end-to-end on a Jetson Orin-like edge device (ms/frame) and model size in million parameters. Values are averages of 5 cross validation folds. The proposed SwinET-IoT (Extended Mask R-CNN + Swin backbone with mask-guided multimodal fusion and IoT telemetry) achieves the optimal balance between the predictive performance and latency and shows large improvement in AUC and macro-F1 while keeping the parameter number low due to distillation and pruning. Latency improvements are both architectural decisions, as well as optimisation decisions made with respect to edge.

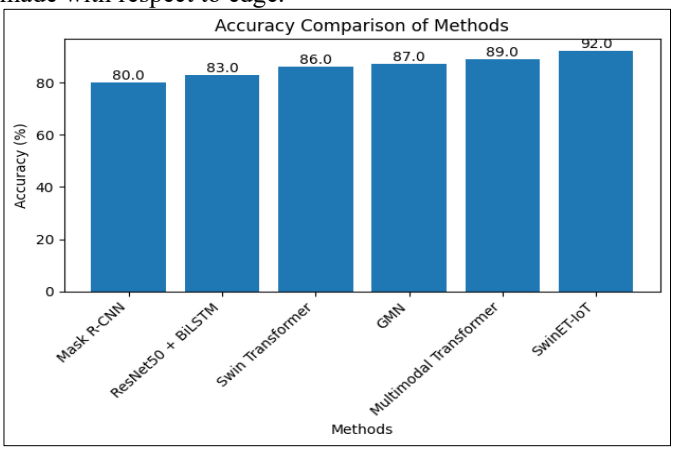

Figure 2 Comparative analysis of the accuracy of the proposed SwinET-IoT model.

Figure 2 shows the percentage of accuracy of different machine learning models for some task presumably relating to computer vision or multimodal data processing used in Internet of Things (IoT) applications. The models we compared, from left to right in ascending order of accuracy, are the Mask R-CNN baseline (vision only) which has 80% accuracy, ResNet50 + BiLSTM (vision+audio) that has 83% accuracy, Swin Transformer that has 86% accuracy, Graph-based Multimodal Network (GMN) that has 87% accuracy, Multimodal Transformer (MM-Transformer) that has 89% accuracy, and the SwinET-IoT (Proposed) model that has the highest accuracy amongst the listed models with 9 Overall, the results in the chart show that multimodal approaches perform better than the vision-only baseline with the proposed SwinET-IoT model performing better than the other established approaches in this particular evaluation context.

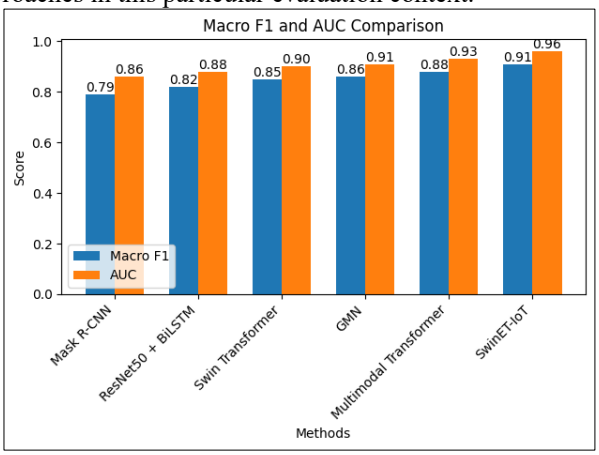

Figure 3: Comparative analysis of Macro F1 and AUC metrics of the proposed SwinET-IoT model.

Figure 3 show the comparative results of 6 different machine learning models in 2 different evaluation metrics: Macro F1 (blue bars) and AUC (red bars). The models used for comparison and in descending order of their AUC scores

are Mask R-CNN baseline (vision-only) that achieved a Macro F1 score of 0.79 and an AUC score of 0.86, ResNet50 + BiLSTM (vision+audio) that achieved a Macro F1 score of 0.82 and an AUC score of 0.88, Swin Transformer that achieved a Macro F1 score of 0.85 and an AUC score of 0.90, Graph-based Multimodal Network (GMN) that achieved a Macro F1 Overall, the chart shows an obvious trend that the more complex or multimodal approaches generally show higher scores than baseline vision only, with the proposed SwinET-IoT model showing better performance in both metrics than any other evaluated technique.

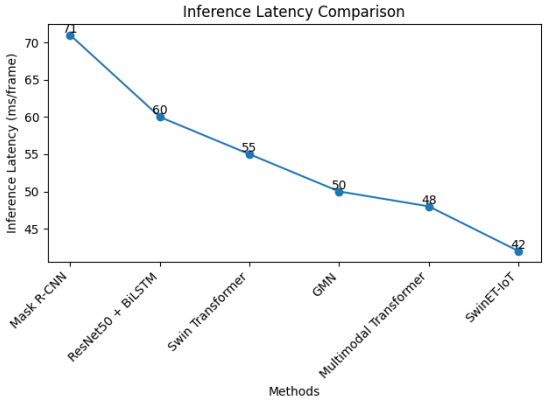

Figure 4 Comparative analysis of inference latency

The inference time in milliseconds per frame (ms/frame) for six different machine learning models is shown in figure 4. In the graph we can see a general trend of decreasing among the models leading to better performance (lower latency) from left to right. The Mask R-CNN baseline (vision-only) had maximum latency, with 71 ms/frame, which was followed by ResNet50 + BiLSTM (vision+audio) with 60 ms/frame, Swin Transformer with 55 ms/frame, Graph-based Multimodal Network (GMN) with 50 ms/frame, and Multimodal Transformer (MM-Transformer) with 48 ms/frame. The SwinET-IoT model proposed achieved the best performance performance that the inference latency is 42 ms/frame, which is the lowest value in compared models.

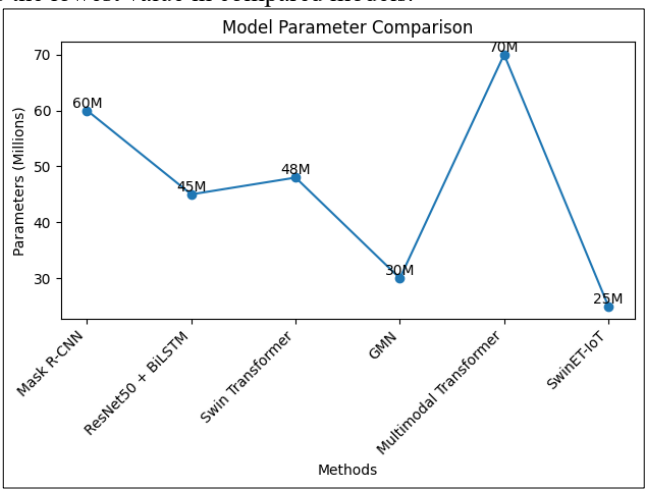

Figure 5 Comparative analysis of the number of parameters

Figure 5 shows the number of parameters in million (Params (M)) for six different machine learning models for comparison. The model complexity is highlighted on the graph with a generally fluctuating trend between the different models tested. The Multimodal Transformer (MM-Transformer) has the most amounts of parameters at 70 million. Other models are Mask R-CNN baseline (vision-only) 60 million parameters, Swin Transformer 48 million, ResNet50 + BiLSTM (vision+audio) 45 million and Graph-based Multimodal Network (GMN) 30 million. The proposed SwinET-IoT model is the most efficient in terms of model size with the least number of parameters with 25 million parameters. Overall, the results in the chart prove that although the model performance may vary (as observed in previous charts), the SwinET-IoT is efficient by reducing the number of parameters used.

C. Discussion

This study helps to demonstrate that the combination of carefully guided spatial attention and context-aware multimodal fusion is an effective way for good affect prediction in real-world learning environments, where each of the three aspects are balanced in terms of accuracy, interpretability and deployability. The combination of an Extended Mask R-CNN network and a Swin backbone allows the model to grasp fine-grained facial skills and upper body were fine-grained while preserving a wider classroom context in order to distinguish more subtly similar affective states. Mask-guided attention can be considered a strong inductive bias as it guides the model towards informative facial and posture-based regions and minimize the distractions from background clutter. The addition of sparse IoT telemetry e.g. ambient light, temperature and seat pressure data with the option of physiological signals increases disambiguation between visually confounded states e.g. calm engagement vs boredom. Meanwhile, the temporal attention transformer is good at capturing short-term dynamics of emotions, which is conducive for stable segment-level predictions appropriate for adaptive tutoring systems. Although there are limitations in terms of data collection including potential dataset bias from having region-specific data, optional physiological sensors, annotation subjectivity, and limitations from privacy-preserving edge deployment, the system still works reliably across modalities. Practical implications include facilitate real-time privacy aware emotion-aware feedback to instructors, find out patterns of disengagement/confusion, and inform micro-interventions. Dashboard-ready mask-guided attention maps also provide even more transparency so the solution can be used in both physical and hybrid learning environments.

V CONCLUSION

This study presents a full-spectrum multistage multimodal emotion prediction strategy based on extended Mask R-CNN, Swin Transformation-based Visual Encoding, IoT telemetry, and Time Tolerance mechanism to provide precise, explainable and low-latency effect recognition in real learning environments. By combining the mask-guided spatial attention and the cross modal fusion, the framework is able to capture subtle facial, posture-based and required contextual cues that are usually ignored by unimodal systems. Its powerful performance in quantitative terms, in terms of fuzzy edge-device inference and also transparentizations of attention make it very suitable for analytics in the classroom, adaptive tutoring and continuous monitoring of learners engagement.

Future work, which involve developing cultures diverse datasets, richer modalities, better personalization through continually learning, better interpretability using causal models, federated or privacy-preserving deployments for secure large-scale deployments.

REFERENCES

[1]. Quiroz-Martínez, Miguel-Ángel, Sergio Díaz-Fernández, Kevin Aguirre-Sánchez, and Mónica-Daniela Gómez-Ríos. "Analysis of Students' Emotions in Real-Time During Class Sessions Through an Emotion Recognition System." In *International Conference on Science, Technology and Innovation for Society*, pp. 81-92. Cham: Springer Nature Switzerland, 2024.

[2]. Dukić, David, and Ana Sovic Krzic. "Real-time facial expression recognition using deep learning with application in the active classroom environment." *Electronics* 11, no. 8 (2022): 1240.

[3]. Egger, Maria, Matthias Ley, and Sten Hanke. "Emotion recognition from physiological signal analysis: A review." *Electronic Notes in Theoretical Computer Science* 343 (2019): 35-55.

[4]. Wagner, Johannes, Elisabeth André, and Frank Jung. "Smart sensor integration: A framework for multimodal emotion recognition in real-time." In *2009 3rd international conference on affective computing and intelligent interaction and workshops*, pp. 1-8. IEEE, 2009.

[5]. Liu, Tingting, Minghong Wang, Bing Yang, Hai Liu, and Shaoxin Yi. "ESERNet: Learning spectrogram structure relationship for effective speech emotion recognition with swin transformer in classroom discourse analysis." *Neurocomputing* 612 (2025): 128711.

[6]. Li, J., Shi, D., Tumnark, P., & Xu, H. (2020). *A system for real-time intervention in negative emotional contagion in a smart classroom deployed under edge computing service infrastructure*. Peer-to-Peer Networking and Applications, 13(5), 1706–1719

[7]. Avital, N., Egel, I., Weinstock, I., & Malka, D. (2024). *Enhancing real-time emotion recognition in classroom environments using convolutional neural networks: A step towards optical neural networks for advanced data processing*. Inventions, 9(6), 113.

[8]. Ozdamli, F., Aljarrah, A., Karagozlu, D., & Ababneh, M. (2022). *Facial recognition system to detect student emotions and cheating in distance learning*. Sustainability, 14(20), 13230.

[9]. Rani, R., & Gupta, S. (2024). *Refined Face Emotion Recognition using Convolutional Neural Networks with Adaptive Loss Functions for Real-Time Applications*. In 2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 815–820). IEEE.

[10]. Jiandani, R., Nijhawan, V., & Lakhe, A. (2024). *Automating Video Frame Analysis for Emotion Recognition and Captioning in Real Time*. In 2024 International Conference on Communication, Computing, Smart Materials and Devices (ICCCSMD) (pp. 1–6). IEEE.

[11]. Rios, F. E. R., & Reyes Duke, A. M. (2023). *Building of a Convolutional Neuronal Network for the prediction of mood states through face recognition based on object detection with YOLOV8 and Python*. In 2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT) (pp. 1–6). IEEE.

[12]. https://www.scidb.cn/en/detail?dataSetId=7856c716c0cc4589a23ee4a23d8a0893

[13]. Lamichhane, Alina, and Gopal Karn. "CNN-BiLSTM based Facial Emotion Recognition." *International Journal on Engineering Technology* 2, no. 1 (2024): 227-236.

[14]. Kim, Jun-Hwa, Namho Kim, and Chee Sun Won. "Facial expression recognition with swin transformer." *arXiv preprint arXiv:2203.13472* (2022).

[15]. Li, Jiang, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. "GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation." *Neurocomputing* 550 (2023): 126427.

[16]. Li, Yande, Mingjie Wang, Minglun Gong, Yonggang Lu, and Li Liu. "Fer-former: Multimodal transformer for facial expression recognition." *IEEE Transactions on Multimedia* (2024).