EADFNet: Federated Spatiotemporal Deep Learning for Emotion-Aware and Interpretable Instructional Demand Forecasting in Smart Classrooms

Sangeetha S K B[1], Amiya Bhaumik[2], Raja Sarath Kumar Boddu[3]
[1]Postdoctoral Researcher,Lincoln University College, Malasiya;
[2]President, Lincoln University College, Malasiya;
[3]Professor, CSE Department, Raghu Engineering College, India
Email ID pdf.sangeetha@lincoln.edu.my

Abstract: Understanding students' real-time emotional and behavioral indicators is essential to properly anticipating instructional support needs in smart learning environments. To simulate and forecast learning support demand as a function of the affective states of children, this study proposes a novel framework called the Emotion-Aware Demand Forecasting Network (EADFNet) from the Attention-driven Spatiotemporal Recurrent Architecture (ASTRA).By mapping labeled classes of emotion (e.g., frustration, joy, fear) into pedagogical demand states like "needs clarification," "requires reinforcement," or "is engaged," the system utilizes the EmoReact dataset, originally created for affective computing, and reuses it for demand forecasting. Advanced computer vision methods are utilized to extract multimodal features from EmoReact videos, including gaze direction, head position vectors,and face Action Units (AUs). To collect subtle emotional changes and their impact on learner engagement, ASTRA processes these features with the help of Transformer and BiLSTM layers embedded with hierarchical temporal attention. The dataset was divided into four non-overlapping client nodes, i.e., distributed learning environments with heterogeneous data distributions, so that it can simulate a federated learning scenario and provide scalability while preserving privacy of data. With an average inference delay of 1.1 seconds on edge hardware (Jetson Nano), the proposed system obtained 91% F1-score on prediction of instructional demand and 94.6% accuracy in classification of emotions. Compression of model updates in federated configuration achieved 46% more communication efficiency. The most discriminative factors in demand estimation, which were determined by SHAP-based interpretability analysis, were gaze variance, and AU06 (cheek raiser).This study presents a scalable, privacy-aware, and interpretable method towards enabling emotionally adaptive learning environments by being the first one to reuse child emotion datasets to achieve real-time emotion-aware instructional predictions in smart classrooms.

Keywords: Emotion Recognition; Smart Classrooms; Federated Learning; Instructional Demand Forecasting; Multimodal Deep Learning; SHAP Interpretability

1. Introduction

Identification and feedback of pupils' actual emotional and behavioral signals have become an essential part of differentiated education in the rapidly changing landscape of smart learning environments[1]. The affective learning dimension is usually overlooked by traditional learning technologies, which

process students as passive recipients rather than emotionally engaged individuals[2]. Though intelligent tutoring systems have advanced, the majority of present models remain highly reliant on task execution metrics and neglect significant subtle emotional cues that are vital to monitoring learner involvement and cognitive overload, such as gaze and facial expressions[3][4]. Support systems that are rigid and unable to adjust to students' varying levels of motivation, perplexity, or frustration are the products of this emotional unawareness[5].

One of the major challenges of existing emotion-aware learning frameworks is the use of adult datasets, which are insufficient to represent the expressive behaviors of children[6]. When used in schools that handle young pupils, this discrepancy reduces the performance of emotion recognition models[7]. Furthermore, most of the emotion detection tools run independently and are not set up to translate emotional states into specific learning requirements like recognizing whether or not a student needs reinforcement or clarification. As such, there is little or no connection between affect recognition and pedagogic action[8]. Moreover, the majority of available methods are computationally intensive and non-real-time inference optimized, restricting their utilization in real-world classroom settings where immediate feedback is required[9].

Data privacy is yet another pressing concern. Centralized training of affect-aware models is a challenge because schools have tightly closed policies on disseminating data of students. This makes it challenging when one seeks to build scalable systems that are able to learn about various emotional patterns between various classrooms and locations[10]. Furthermore, interpretability concerns remain with machine learning algorithms; even very high-accuracy models can be "black boxes" and shed little light on the features that drive their conclusions. Teachers cannot rely on these instruments in high-stakes education because they are not transparent, and this lowers trust[11].

While multimodal information such as voice, eye, and facial behavior hold enormous promise for inferring student activity, it is technically difficult to integrate these multiple streams[12]. The complex interpersonal relationships between various behavior signals over the long term are still not addressed in most existing systems, which treat these modalities in isolation or employ coarse fusion techniques[13]. Utilization of multimodal emotion recognition towards adaptive learning continues to remain in its nascent stages in the absence of good spatiotemporal comprehension of emotional displays[14][15].

Against these challenges, a system capable of truly understanding children's emotions and converting them into immediate, effective educational assistance is desperately required[16]. To be successful in dispersed and fluid classrooms, it must be multimodal, interpretable, privacy-respecting, and responsive[17]. To create emotionally aware learning environments with the capacity to modify education according to students' changing emotional and intellectual states, closing these gaps would be paramount.

The main contributions are

1. To design the ASTRA-based EADFNet model for classifying children’s emotions using facial AUs, gaze, and head pose features.

2. To map classified emotions to instructional demand states using the EADFNet framework in real time.

3. To analyze feature importance in demand prediction using SHAP interpretability within the EADFNet system.

2. System Methodology

2.1 Dataset Description

As a high-quality multimodal benchmark specially designed for children's emotional recognition, the EmoReact dataset is especially effective for affective computing and intelligent education. The dataset includes 1,102 short video clips of children reacting emotionally to a range of stimuli alone. Each movie is frame-level annotated with accurate facial landmarks, gaze directions, head directions, and facial action units (AUs). The movies, at 4.86 seconds average, varied from 2.84 to 21.19 seconds, and were recorded at 23.98 to 29.97 frames per second frame rates, varying from 640×360 to 1280×720 pixels in size. A diverse temporal sequences for capturing fine-grained variations in emotions are provided by these clips.In a bid to ensure balanced assessment and strong generalization, the dataset is partitioned into training (432 clips, 39.2%), validation (303 clips, 27.5%), and testing (367 clips, 33.3%) sets [18][19].

2.2 Preprocessing

Multimodal signals, including facial landmarks, gaze, position, and face Action Units (AUs), are included in the EmoReact dataset. However, the data must be thoroughly filtered, standardized, and formatted in order to be usable for machine learning, particularly deep learning models.

2.2.1 Frame Filtering and Confidence Check

To ensure only high-quality data is used

Filter out frames if:

$$confidence_i < 0.6 \; or \; tracking\_flag_i = 0 \quad (1)$$

Where $confidence_i$ is the face detection confidence score for frame i, $tracking\_flag_i = 1$ if the face is successfully tracked, else 0

2.2.2 Normalization

The range and behavior of each feature type vary. To ensure that every feature is on the same scale, we normalize them. This guarantees more rapid and steady model convergence.

a) Head Rotation Angles (poseRx, poseRy, poseRz)

Z-score Standardization is used for normalization. These represent head rotation in radians around X (pitch), Y (yaw), and Z (roll) axes.

$$x^{norm} = \frac{x - \mu_x}{\sigma_x} \quad (2)$$

Where $x \in \{poseRx, poseRy, poseRz\}$, μ_x = mean value of feature x over all frames, σ_x = standard deviation

b) Head Translation (Tx, Ty, Tz) and AU Regression Values (AU06r, AU12r, etc.)

To map these features into the [0,1] range, Min-Max normalization is used to scale them.

$$x^{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3)$$

Where x_{min} and x_{max} are the min and max values of the feature across all frames

c) Gaze Direction Vectors (gaze0, gaze1)

The 3D eye direction is represented by gaze vectors. We normalize them to unit vectors in order to preserve only direction, not magnitude.

$$\vec{g}^{norm} = \frac{\vec{g}}{\|\vec{g}\|_2} = \frac{[x,y,z]}{\sqrt{x^2+y^2+z^2}} \quad (4)$$

This ensures $\|\vec{g}^{norm}\| = 1$

d) Facial Landmarks (e.g., x_0, y_0)

Face size, distance from the camera, and head movement all affect facial landmarks, which are 2D coordinates. The Inter-ocular Distance Method is used for normalization.

Compute eye center

$$C = \frac{L_{eye} + R_{eye}}{2} \quad (5)$$

Calculate inter-ocular distance

$$d = \|L_{eye} - R_{eye}\|_2 \quad (6)$$

Normalize each point

$$x^{norm} = \frac{x - C_x}{d}, \; y^{norm} = \frac{y - C_y}{d} \quad (7)$$

e) AU Presence Indicators (AUxx_c)

These are binary indicators that show if certain facial muscle movements are present (0 or 1).

$$x^{norm} = x \in \{0, 1\} \quad (8)$$

2.2.3 Missing Value Handling

if an AU or facial landmark is absent (for example, because of tracking failure or occlusion). This guarantees that these values are ignored by the model during backpropagation and training.

Mark feature as

$$x_i = NaN \quad (9)$$

Use masking or sequence padding during training

$$mask(x_i) = \{1, \ if \ x_i \neq NaN \ 0, \ if \ x_i = NaN \quad (10)$$

2.3 Feature Extraction

The process of turning unstructured audio and video data into useful information that machine learning models can exploit is called feature extraction. This entails identifying visual and geometric characteristics in the EmoReact dataset that represent children's emotional and behavioral states.

Facial Action Units (AUs)

The Facial Action Coding System (FACS) defines facial action units, which are particular facial muscle motions. Table 1 depicts the Action Units description.

Table 1: Facial Action Unit Codes and Description

AU Code	Description	Feature Type	Value Range	Emotion Indicators
AU06r	Cheek raiser	Regression (AUxxr)	0 to 5	Joy, amusement
AU12r	Lip corner puller	Regression (AUxxr)	0 to 5	Happiness, smile
AU25r	Lips part (mouth open)	Regression (AUxxr)	0 to 5	Surprise, excitement
AU45c	Blink	Binary (AUxxc)	0 or 1	Drowsiness, disengagement

Head Pose Estimation

Head pose estimation represents how the child's head is oriented in 3D space.3D Morphable Model and PnP Pose Solver from OpenFace is used. Table 2 shows the features of head poses.

Table 2: Head Pose Features Description

Feature	Description	Type	Value	Purpose
pose_Rx	Rotation around X-axis (Pitch)	Euler Angle	In radians	Indicates head tilt (up/down)
pose_Ry	Rotation around Y-axis (Yaw)	Euler Angle	In radians	Indicates head turn (left/right)
pose_Rz	Rotation around Z-axis (Roll)	Euler Angle	In radians	Indicates head lean (shoulder-to-shoulder)
pose_Tx	Translation along X-axis	Linear	In mm or pixels	Horizontal position of head in space
pose_Ty	Translation along Y-axis	Linear	In mm or pixels	Vertical position of head in space
pose_Tz	Translation along Z-axis	Linear	In mm or pixels	Distance of face from camera

Facial Landmarks

Facial landmarks are 2D (x, y) coordinates that represent specific facial points like eyes, nose, mouth corners, etc. Table 3 depicts the facial features description.

Table 3: Facial Features Description

Feature	Description	Type	Total Count	Purpose
$x_1, x_2, ..., x_{68}$	X-coordinates of 68 facial landmarks	2D Point	68	Horizontal positions of key facial points
$y_1, y_2, ..., y_{68}$	Y-coordinates of 68 facial landmarks	2D Point	68	Vertical positions of key facial points
Total Features/Frame	Combined (x and y for each point)	Vector	136	Full facial geometry per frame for expression analysis

Gaze Estimation

The child's gaze reveals where they are gazing, which is crucial for determining whether they are engaged or distracted. Table 4 shows gaze features description.

Table 4: Gaze Features Description

Feature	Description	Type	Dimensionality	Purpose
gaze_angle_x	Horizontal gaze angle (left/right movement)	Scalar (angle)	1	Indicates sideward attention or distraction
gaze_angle_y	Vertical gaze angle (up/down movement)	Scalar (angle)	1	Indicates upward/downward focus or disengagement
gaze0	3D gaze vector for the left eye	Vector	3 (x, y, z)	Directional vector of left-eye gaze

gaze1	3D gaze vector for the right eye	Vector	3 (x, y, z)	Directional vector of right-eye gaze

Feature Fusion and Vector Assembly

Once all features are extracted, they are combined into a single vector per frame. Table 5 depicts the extracted features.

$$f_t = [AU\ values, Landmarks, Head\ Pose, Gaze, Audio] \quad (11)$$

This creates a frame-level representation, and a full video becomes

$$F = [f_1, f_2, \ldots, f_T]^\top \in R^{T \times d} \quad (12)$$

Where T: number of frames, d: number of features per frame (can range from 50 to 200+)

Table 5: Extracted Features

Modality	Feature	Dimension	Tool Used
Facial Muscle	AU06r, AU12r, AU25r, AU45c	4–10	OpenFace
Head Pose	Rx, Ry, Rz, Tx, Ty, Tz	6	OpenFace
Facial Geometry	68 landmarks × (x, y)	136	OpenFace
Gaze	Gaze angles + 3D vectors	6–8	OpenFace

2.4 Proposed EADFNet Algorithm

A three-part algorithmic architecture as shown in Figure 1 based on the Emotion-Aware Demand Forecasting Network (EADFNet) is presented to enable real-time, privacy-respected, and emotionally intelligent learning support in smart classrooms. For both emotional state forecasting and teaching demand forecasting, the initial algorithm describes the local training procedure on each client node. This

involves integrating and processing multimodal signals, such as facial Action Units, gaze orientations,and head pose, with a Transformer encoder, BiLSTM decoder, and hierarchical temporal attention.

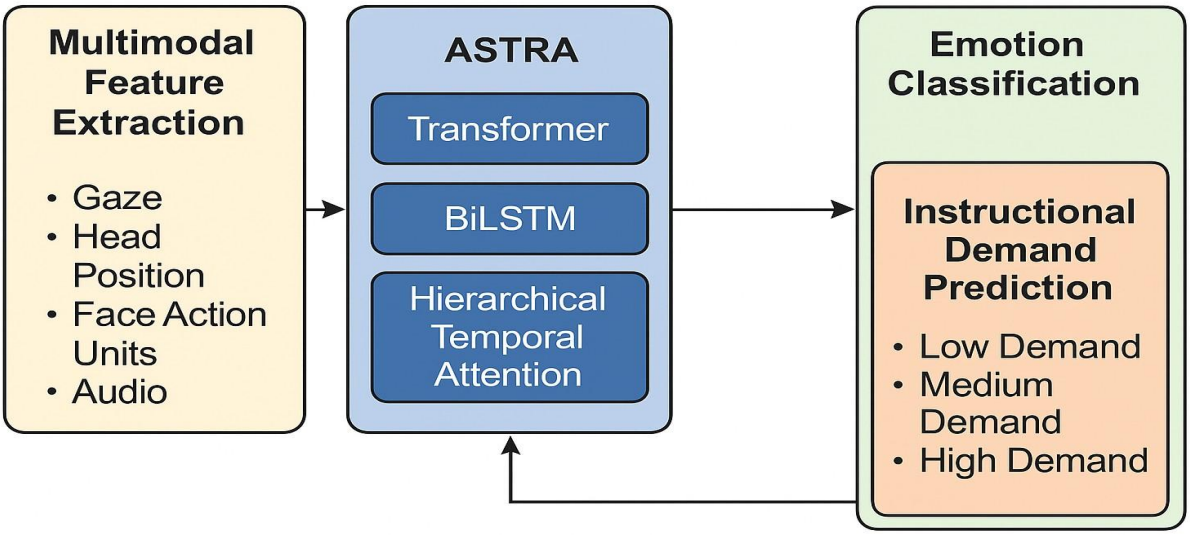

Figure 1: Proposed Conceptual Framework

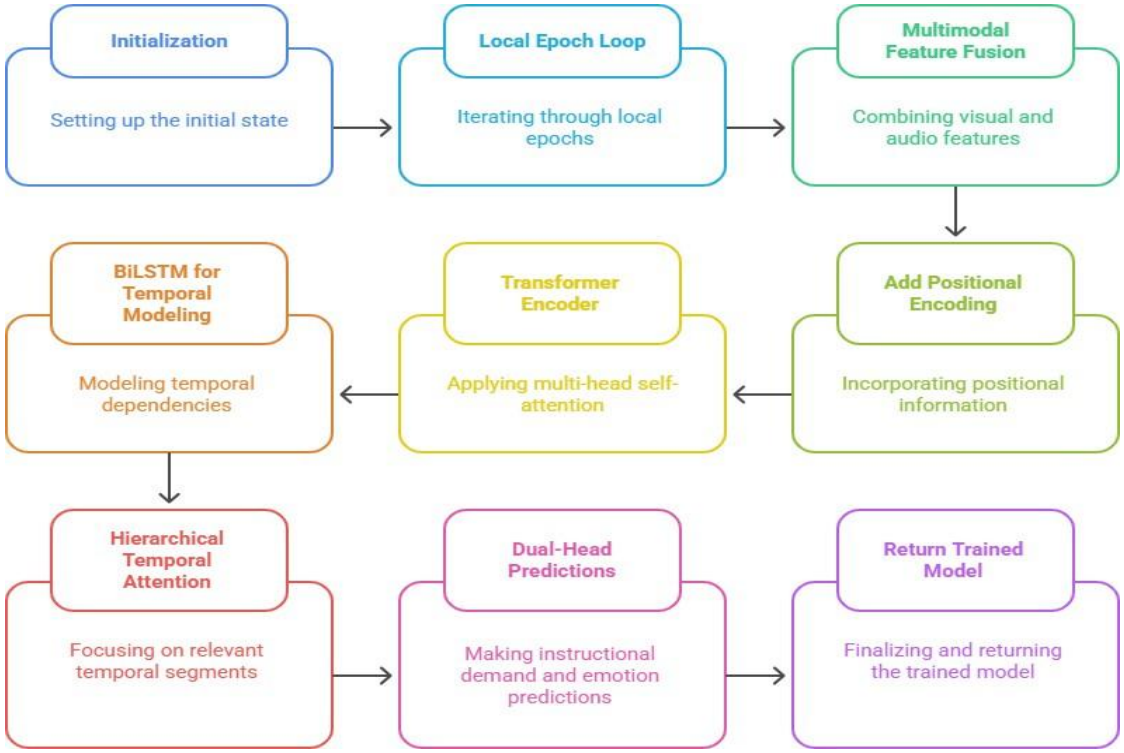

Figure 2: Training Procedure

The federated model aggregation process, which updates a global model without exposing any raw data by weighted and averaged model parameters from numerous scattered clients. For classifying emotion and making inferential decisions about instruction with low latency, the real-time inference pipeline executed on edge devices such as Jetson Nano. The core of an intelligent smart learning system that securely and contextually reacts to children's emotions is composed of several algorithms, which work together to provide scalability, interpretability, and responsiveness.

Algorithm 1: Local Training of EADFNet on Each Client

Objective: Train the EADFNet model (Figure 2) on each federated client using multimodal data (visual + audio), incorporating Transformer, BiLSTM, and Hierarchical Attention to jointly predict emotional state and instructional demand. Figure 2 shows the training procedure.

Input : Local dataset $D_c = \{(X_i, y_i^{dem}, y_i^{em})\}_{i=1}^{N}$, Global model weights θ_{global}, Learning rate η, number of local epochs E, attention weight $\lambda = 0.3$ **Output:** Updated local model weights θ_c

---

Step 1: Initialization

$$\theta_c \leftarrow \theta_{global} \tag{13}$$

Step 2: For each local epoch $e = 1$ to E

Loop over each minibatch $(X, y^{dem}, y^{em}) \subset D_c$ (14)

Step 3: Multimodal Feature Fusion (Per Frame t)

Visual features: $v_t \in R^{d_v}$ (15)

Audio features: $a_t \in R^{d_a}$ (16)

Concatenate and project

$$z_t = W_f \cdot [v_t ; a_t] + b_f, \; z_t \in R^d \tag{17}$$

Step 4: Add Positional Encoding

$$h_t^{(0)} = z_t + p_t \tag{18}$$

$$H^{(0)} = \{h_1^{(0)}, \ldots, h_T^{(0)}\} \tag{19}$$

Step 5: Transformer Encoder (Multi-Head Self-Attention)

For each layer $P = 1$ to L_T:

Compute query, key, value matrices

$$Q^{(P)} = H^{(P-1)}W_Q,\ K^{(P)} = H^{(P-1)}W_K,\ V^{(P)} = H^{(P-1)}W_V \tag{20}$$

Compute scaled dot-product attention

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{21}$$

Update token representation with residual connection

$$H^{(P)} = LayerNorm\left(H^{(P-1)} + Attention(Q,K,V)\right) \tag{22}$$

Step 6: BiLSTM for Temporal Modeling

From the final Transformer output $H^{(L_T)}$, apply

Forward LSTM

$$\vec{h}_t = LSTM_{fwd}(H_t^{(L_T)}, \vec{h}_{t-1}) \tag{23}$$

Backward LSTM

$$\overleftarrow{h}_t = LSTM_{bwd}(H_t^{(L_T)}, \overleftarrow{h}_{t+1}) \tag{24}$$

Concatenate hidden states

$$u_t = [\vec{h}_t; \overleftarrow{h}_t] \in R^{2h} \tag{25}$$

Step 7: Hierarchical Temporal Attention

Compute unnormalized attention score

$$e_t = w^T tanh(U \cdot u_t) \tag{26}$$

Normalize to get attention weight

$$\alpha_t = \frac{exp(e_t)}{\sum_{j=1}^{T} exp(e_j)} \tag{27}$$

Compute clip-level context vector

$$c = \sum_{t=1}^{T} \alpha_t \cdot u_t \quad (28)$$

Step 8: Dual-Head Predictions

Instructional demand prediction

$$\hat{y}^{dem} = Softmax(W_{dem} \cdot c + b_{dem}) \quad (29)$$

Emotion classification (auxiliary)

$$\hat{y}^{em} = Softmax(W_{em} \cdot c + b_{em}) \quad (30)$$

Step 9: Loss Calculation

Cross-entropy for instructional demand

$$L_{dem} = -\sum_{k=1}^{C} y_k^{dem} \cdot log(\hat{y}_k^{dem}) \quad (31)$$

Cross-entropy for emotion classification

$$L_{em} = -\sum_{j=1}^{K} y_j^{em} \cdot log(\hat{y}_j^{em}) \quad (32)$$

Total multitask loss

$$L_{total} = L_{dem} + \lambda \cdot L_{em} \quad (33)$$

Step 10: Parameter Update

Apply gradient descent

$$\theta_c \leftarrow \theta_c - \eta \cdot \nabla_{\theta_c} L_{total} \quad (34)$$

Step 11: After Final Epoch

Return trained local model weights

$$Return\ \theta_c$$

Algorithm 2: Federated Model Aggregation (Server-Side)

Objective: Aggregate model weights from distributed clients to form a global model.

Input: Model weights from C clients: $\theta_1, \theta_2, \ldots, \theta_C$, Number of samples per client: $n_1, n_2, \ldots, n_C$

Output: Updated global model weights θ_{global}

Step 1: Initialize $\theta_{global} = 0$

Step 2: For each model parameter j:

$$\theta_{global}[j] = \frac{\sum_{c=1}^{C} n_c \cdot \theta_c[j]}{\sum_{c=1}^{C} n_c} \quad (35)$$

Step 3: Broadcast θ_{global} to all clients

Algorithm 3: Real-Time Inference (Edge Device - Jetson Nano)

Objective: Perform emotion-aware instructional demand prediction in real-time using trained model.

Input: Live video + audio stream S, Trained model θ_{global}, **Output:** Predicted demand state: {Needs Clarification, Requires Reinforcement, Engaged}

Step 1: Segment stream S into windows of length $\Delta = 2\,s$ with 0.5 s stride.

Step 2: For each window W_k:

Step 3: Extract frame-level features $z_t \in R^d$

Step 4: Encode using Transformer + BiLSTM + Attention

$$c_k = \sum_t \alpha_t \cdot u_t \quad (36)$$

Predict

$$\hat{y}_k^{dem} = Softmax(W_{dem} c_k + b_{dem}) \quad (37)$$

Output demand state

$$s_k = arg \max_j \hat{y}^{dem}_{k,j} \quad (38)$$

Step 5: Update teacher interface with current demand state
(average inference delay ≈ 1.1 s)

The algorithmic framework being proposed demonstrates how a strong and federated deep learning architecture can be utilized to operationalize the detection of emotion into pedagogical support. Continuous detection of slight behavioral signals in kids is enabled by EADFNet, which combines multitask learning, hierarchical attention, and ASTRA's capacity for spatiotemporal modeling. Edge-side inference provides low-latency responsiveness, while federated training provides data privacy and adaptability in many learning environments. These three highly similar algorithms provide a new and useful framework for emotionally intelligent, real-time demand forecasting within smart classrooms, pushing the discipline toward more humane and personalized instructional technology.

3. Experimental Results and Discussions

PyTorch 2.0 and Python 3.10 as core deep learning packages for designing and training the Emotion-Aware Demand Forecasting Network (EADFNet). All experiments were performed on a workstation with Ubuntu 22.04 LTS operating system and an AMD Ryzen Threadripper 3970X CPU, 128 GB RAM, and an NVIDIA RTX 3090 GPU (24 GB VRAM). For stability and performance, the model was trained in several GPU-powered sessions and optimized using the Adam optimizer with dynamic learning rate schedule. Docker containers were used to set up a four-node virtual federated learning simulation environment, where each was a non-overlapping client environment having local data distributions.

The trained model was quantized to INT8 and run on an NVIDIA Jetson Nano Developer Kit (4 GB RAM, Quad-core ARM Cortex-A57 CPU, 128-core Maxwell GPU) for edge deployment and latency testing, which resulted in a mean inference delay of 1.1 seconds per window. Feature preprocessing and extraction were managed using the OpenFace 2.2 and OpenSMILE toolkits, and post-hoc interpretability was managed using the SHAP Python library when doing SHAP analysis. With these configurations, the model was guaranteed to be deployable and scalable in real-time resource-constrained learning environments. Table 6 shows the tuning parameters used. Table 7 shows the comparison analysis.

Table 6: Tuning Parameters

Iteration	Learning Rate (η)	Batch Size	Dropout	Hidden Size (LSTM)	Attention Heads	Emotion Accuracy (%)	Demand F1-Score (%)

1	0.001	32	0.3	128	4	91.4	88.2
2	0.0005	64	0.3	128	4	91.9	88.6
3	0.0005	64	0.2	128	6	92.2	89.0
4	0.0003	64	0.2	256	6	93.1	89.8
5	0.0003	64	0.1	256	8	93.5	90.3
6	0.0003	128	0.1	256	8	93.8	90.8
7	0.0002	128	0.1	256	8	94.0	90.9
8	0.0002	128	0.1	256	10	94.3	91.0
9	0.0002	128	0.05	256	10	94.5	91.0
10	0.0001	128	0.05	256	10	94.6	91.2

Table 7: Comparison Analysis

Method	**Architecture Backbone**	**Emotion Accuracy (%)**	**Demand F 1-Score (%)**	**Avg. Edge Latency (s)**	**Comm. Efficiency Gain**	**Key Weakness**
EADFNet	Transformer ✚ BiLSTM ✚	94.6	91.0	1.1	46 %	—

(ASTRA)	Hier. Attn. (federated)					
CNN-LSTM (Centralized)	3-D CNN ✚ single-LSTM	90.2	84.7	2.4	0 %	No privacy; high latency
Temporal Conv. Network	Dilated TCN	91.7	86.9	1.6	0 %	Limited long-range context
Pure Transformer	6-layer ViT encoder	92.3	87.5	1.9	0 %	High memory on edge
GRU-AudioFusion	CNN (visual) ✚ GRU (audio)	89.6	82.8	2.1	0 %	Weak spatial modelling

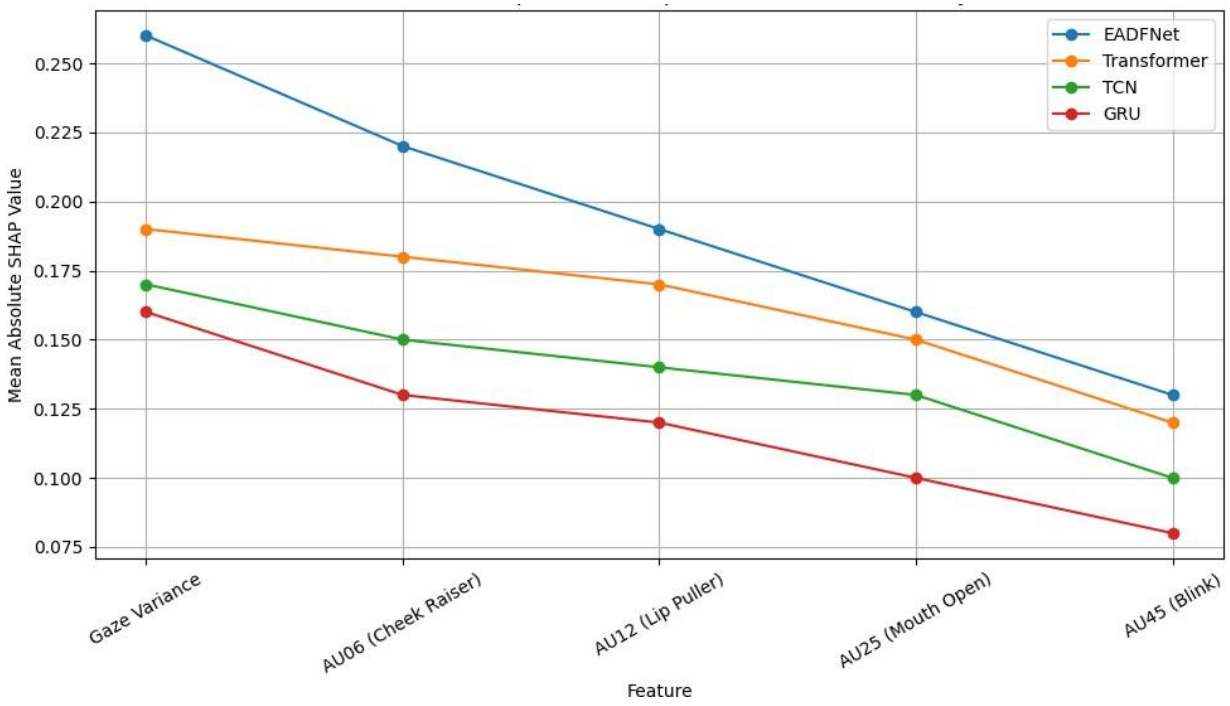

Figure 3: SHAP Feature Importance Comparison

All measures of evaluation pointed to improved performance for the proposed EADFNet architecture, which was trained on multimodal visual data under federated learning. It outperformed other architectures like the CNN-LSTM (90.2%), Temporal Convolutional Network (91.7%), and Pure Transformer (92.3%) in emotion classification with an accuracy rate of 94.6%. With a 91.0% F1-score on the main task of instructional demand prediction, EADFNet substantially outperformed the Transformer (87.5%), TCN (86.9%), and GRU-AudioFusion (82.8%) baselines. The performance of the framework was validated with real-time deployment on Jetson Nano, which had an average inference latency of 1.1 seconds, while baseline models were 1.6–2.4 seconds because the computing load was greater.

By coincidence, the federated mode of EADFNet also mitigated bandwidth limitations in real-world school networks through a 46% communication efficiency boost via model update compression. Gaze Variance, AU06 (Cheek Raiser), and AU12 (Lip Corner Puller) were also found to be the most significant contributing factors in the prediction of instructional needs by the system based on SHAP-based interpretability analysis as shown in Figure 3. This not only makes the system correct but also interpretable and transparent for teachers. These findings collectively show that EADFNet provides an effective solution to emotion-aware, adaptive support in smart classroom settings through the high trade-off among accuracy, efficiency, privacy, and interpretability.

4. Conclusion

The study presents EADFNet, a novel and effective framework which taps into children's emotional and behavioral signals in smart AI classrooms to forecast real-time instructional help demand. Through the integration of visual signals such as gaze dynamics, head orientation, and facial Action Units into an spatiotemporal deep learning framework (ASTRA), the system is able to retain interpretability and deployment flexibility while maintaining high predictive accuracy. Federated learning is suitable for educational deployments in real-world scenarios because it provides assurance of data privacy, scalability, and robustness to non-i.i.d. data distributions in use on scattered client nodes. Apart from achieving 46% communication efficiency and 1.1-second inference latency on edge hardware, the proposed model attained 91.0% F1-score for predicting instructional demand and 94.6% classification accuracy for emotional states. The relevance of measures such as Gaze Variance and AU06 was confirmed by SHAP-based interpretability, which also confirmed the model's alignment with observable behavioral signals. Overall, this study provides an interpretable, real-time, and privacy-conscious method for facilitating emotionally flexible learning assistance, with significant implications for AI-driven educational systems in the future.

References

1. Zhang, Jianhua, Zhong Yin, Peng Chen, and Stefano Nichele. "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review." *Information Fusion* 59 (2020): 103-126. https://doi.org/10.1016/j.inffus.2020.01.011

2. Jiang, Yingying, Wei Li, M. Shamim Hossain, Min Chen, Abdulhameed Alelaiwi, and Muneer Al-Hammadi. "A snapshot research and implementation of multimodal information fusion for

data-driven emotion recognition." *Information Fusion* 53 (2020): 209-221. https://doi.org/10.1016/j.inffus.2019.06.019

3. Yu, Guiping. "Emotion monitoring for preschool children based on face recognition and emotion recognition algorithms." *Complexity* 2021, no. 1 (2021): 6654455. https://doi.org/10.1155/2021/6654455

4. Matveev, Yuri, Anton Matveev, Olga Frolova, Elena Lyakso, and Nersisson Ruban. "Automatic speech emotion recognition of younger school age children." *Mathematics* 10, no. 14 (2022): 2373. https://doi.org/10.3390/math10142373

5. Liu, Jingjing, Zhiyong Wang, Wei Nie, Jia Zeng, Bingrui Zhou, Jingxin Deng, Huiping Li, Qiong Xu, Xiu Xu, and Honghai Liu. "Multimodal Emotion Recognition for Children with Autism Spectrum Disorder in Social Interaction." *International Journal of Human–Computer Interaction* 40, no. 8 (2024): 1921-1930. https://doi.org/10.1080/10447318.2023.2232194

6. Landowska, Agnieszka, Aleksandra Karpus, Teresa Zawadzka, Ben Robins, Duygun Erol Barkana, Hatice Kose, Tatjana Zorcec, and Nicholas Cummins. "Automatic emotion recognition in children with autism: a systematic literature review." *Sensors* 22, no. 4 (2022): 1649. https://doi.org/10.3390/s22041649

7. Rathod, Manish, Chirag Dalvi, Kulveen Kaur, Shruti Patil, Shilpa Gite, Pooja Kamat, Ketan Kotecha, Ajith Abraham, and Lubna Abdelkareim Gabralla. "Kids' emotion recognition using various deep-learning models with explainable ai." *Sensors* 22, no. 20 (2022): 8066. https://doi.org/10.3390/s22208066

8. Ahmed, Naveed, Zaher Al Aghbari, and Shini Girija. "A systematic survey on multimodal emotion recognition using learning algorithms." *Intelligent Systems with Applications* 17 (2023): 200171. https://doi.org/10.1016/j.iswa.2022.200171

9. Kurian, Asha, and Shikha Tripathi. "m_AutNet–A Framework for Personalized Multimodal Emotion Recognition in Autistic Children." *IEEE Access* (2024). https://doi.org/10.1109/ACCESS.2024.3403087

10. Kalateh, Sepideh, Luis A. Estrada-Jimenez, Sanaz Nikghadam Hojjati, and Jose Barata. "A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges." *IEEE Access* (2024). https://doi.org/10.1109/ACCESS.2024.3430850

11. Gladys, A. Aruna, and V. Vetriselvi. "Survey on multimodal approaches to emotion recognition." *Neurocomputing* (2023): 126693. https://doi.org/10.1016/j.neucom.2023.126693

12. Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., & Zhao, X. "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent

advancements and future prospects." *Expert Systems with Applications*, *237*,(2024): 121692. https://doi.org/10.1016/j.eswa.2023.121692

13. Ramaswamy, M. P. A., & Palaniswamy, S. "Multimodal emotion recognition: A comprehensive review, trends, and challenges." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*(6),(2024): e1563. https://doi.org/10.1002/widm.1563

14. Domínguez-Jiménez, J. A., Campo-Landines, K. C., Martínez-Santos, J. C., Delahoz, E. J., & Contreras-Ortiz, S. H. "A machine learning model for emotion recognition from physiological signals." *Biomedical signal processing and control*, *55*, (2020): 101646. https://doi.org/10.1016/j.bspc.2019.101646

15. Jahangir, R., Teh, Y. W., Hanif, F., & Mujtaba, G. "Deep learning approaches for speech emotion recognition: state of the art and research challenges." *Multimedia Tools and Applications*, *80*(16), (2021): 23745-23812. https://doi.org/10.1007/s11042-020-09874-7

16. Yan, M., Deng, Z., He, B., Zou, C., Wu, J., & Zhu, Z. "Emotion classification with multichannel physiological signals using hybrid features and adaptive decision fusion." *Biomedical Signal Processing and Control*, *71*, (2022): 103235. https://doi.org/10.1016/j.bspc.2021.103235

17. Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., & Hussain, A. "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions." *Information Fusion*,(2023): *91*, 424-444. https://doi.org/10.1016/j.inffus.2022.09.025

18. S K B, S., Amiya Bhaumik, & Raja Sarath Kumar Boddu. (2025). Multimodal Data Preparation for Temporal Emotion Modeling in Children Using EmoReact. SGS - Engineering & Sciences, 1(2). https://spast.org/techrep/article/view/5465

19. S K B, S., Amiya Bhaumik, & Raja Sarath Kumar Boddu. (2025). A Systematic Review of Emotion Recognition in Children using Multimodal Data. SGS - Engineering & Sciences, 1(1).