A Multi-Modal Deep Learning Framework for Student Soft Skills Development in Adaptive Learning Environments

Marzhan Bekbolat[1], Kamalbek Berkimbayev[2], Rustam Abdrakhmanov[3],
Serik Kenesbayev[4], Nuraim Ibragimova[5], Zhaksylyk Dzhanabayev[6]
Khoja Akhmet Yassawi International Kazakh-Turkish University, Kazakhstan[1, 2]
International University of Tourism and Hospitality, Turkistan, Kazakhstan[3]
Kazakh National Women's Teacher Training University, Almaty, Kazakhstan[4]
Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan[5]
Mukhtar Auezov South Kazakhstan Research University, Shymkent, Kazakhstan[6]

***Abstract*—The rapid evolution of artificial intelligence and adaptive educational technologies has created increasing demand for intelligent systems capable of automatically assessing and enhancing student soft skills within digital learning environments. This study proposes MST-SoftNet, a multimodal transformer-based deep learning framework designed for adaptive soft skill assessment and personalized educational recommendation. The proposed architecture integrates heterogeneous educational modalities, including textual interactions, speech signals, facial expressions, behavioral engagement patterns, performance indicators, and feedback information, within a unified hierarchical transformer fusion framework. Modality-specific encoders, cross-modal attention mechanisms, explainable attention visualization modules, and adaptive recommendation components were incorporated to improve both predictive performance and interpretability. Experimental evaluation was conducted using multiple baseline deep learning architectures, including CNN, LSTM, Transformer, and multimodal CNN-LSTM models. The proposed MST-SoftNet framework achieved superior performance across all evaluation metrics, attaining 93.67% accuracy, 92.11% F1-score, and 96.05% AUC, while simultaneously demonstrating reduced inference latency and improved computational efficiency. Attention visualization analysis further confirmed the capability of the framework to identify semantically meaningful multimodal behavioral patterns associated with communication, collaboration, leadership, creativity, emotional intelligence, and self-regulation competencies. Longitudinal adaptive learning experiments additionally demonstrated substantial improvement in student soft skill progression over time. The obtained results indicate that MST-SoftNet establishes a robust, interpretable, and scalable foundation for next-generation intelligent educational systems focused on personalized soft skill development and adaptive learning optimization.**

Keywords*—Deep learning; transformer architecture; soft skill assessment; adaptive learning environments; personalized learning; student behavior analysis; multimodal fusion

I. Introduction

The rapid advancement of artificial intelligence and educational technologies has substantially transformed the structure and dynamics of modern learning environments. Beyond traditional academic achievement, increasing attention is now directed toward the development of soft skills, including communication, collaboration, leadership, creativity, emotional intelligence, and critical thinking, which are considered fundamental competencies for success in the twenty-first century workforce and digital society [1]. Educational institutions are progressively integrating adaptive learning systems and intelligent tutoring platforms to support individualized instruction and learner-centered pedagogical strategies [2]. However, despite remarkable progress in adaptive educational technologies, most existing systems remain predominantly focused on cognitive performance indicators such as grades, quizzes, and task completion, while the automatic assessment and development of soft skills continue to represent a major research challenge [3].

Recent advances in deep learning, multimodal representation learning, and transformer-based architectures have opened new opportunities for analyzing complex behavioral and affective patterns within educational environments [4]. Multimodal learning frameworks are capable of simultaneously processing heterogeneous data sources, including textual interactions, speech signals, facial expressions, behavioral logs, and performance analytics, thereby enabling a more comprehensive understanding of student learning behavior and interpersonal competencies [5]. Transformer architectures, particularly those employing self-attention and cross-modal attention mechanisms, have demonstrated exceptional capability in capturing long-range dependencies and contextual relationships across multiple modalities [6]. Consequently, multimodal transformer networks have emerged as promising solutions for intelligent educational analytics and personalized learning systems.

Despite these technological advancements, several limitations remain insufficiently addressed in existing studies. Many current approaches rely on unimodal data representations or shallow machine learning techniques that fail to capture the multidimensional nature of soft skill development [7]. Furthermore, existing systems often lack explainability, adaptive recommendation mechanisms, and longitudinal

monitoring capabilities necessary for real-world educational deployment [8]. The complexity of modeling human-centered soft skills additionally requires architectures capable of integrating temporal behavioral dynamics, emotional states, collaborative interactions, and semantic communication patterns within a unified learning framework. Therefore, there exists a critical need for robust multimodal architectures capable of accurately assessing and continuously improving soft skills in adaptive educational ecosystems.

To address these challenges, this study proposes MST-SoftNet, a multimodal transformer-based deep learning framework for student soft skills development in adaptive learning environments. The proposed architecture integrates modality-specific encoders, hierarchical cross-modal transformer fusion, attention-based explainability modules, and adaptive recommendation mechanisms to provide comprehensive soft skill assessment and personalized intervention strategies [9]. The framework leverages multimodal educational data to generate discriminative soft skill representations while simultaneously supporting real-time analytics and adaptive feedback generation. Experimental evaluation demonstrates the effectiveness of the proposed model in achieving improved classification performance, enhanced interpretability, and robust adaptive learning capabilities across multiple soft skill dimensions.

II. Related Works

The integration of artificial intelligence into educational systems has significantly accelerated the development of intelligent adaptive learning environments capable of personalizing educational experiences and improving student engagement [10]. Early adaptive learning systems primarily relied on rule-based recommendation mechanisms and statistical learner models, which demonstrated limited capability in capturing the complexity of human behavior and interpersonal competencies [11]. With the emergence of machine learning techniques, researchers began incorporating predictive analytics and data-driven personalization approaches to improve educational outcomes and learner adaptability [12]. Nevertheless, many of these systems remained predominantly focused on academic performance metrics rather than holistic student development, particularly soft skills and behavioral competencies.

Deep learning approaches have recently demonstrated substantial effectiveness in educational data mining and intelligent tutoring applications due to their capability to automatically learn hierarchical representations from high-dimensional data [13]. Convolutional neural networks and recurrent neural networks have been widely utilized for student engagement detection, behavioral analysis, and educational content recommendation [14]. CNN-based architectures have shown strong performance in extracting visual and spatial features from facial expressions and classroom activity recordings, while LSTM and GRU models have proven effective in modeling temporal learning patterns and sequential student interactions [15].

However, traditional sequential architectures often encounter limitations in capturing long-range contextual dependencies and complex multimodal interactions across heterogeneous educational data streams [16].

Transformer-based architectures have emerged as highly promising solutions for educational intelligence systems due to their self-attention mechanisms and parallel representation learning capability [17]. Vision Transformers and multimodal transformers have recently been employed for emotion recognition, engagement prediction, and adaptive recommendation systems in digital education platforms [18]. Cross-modal attention mechanisms enable these architectures to integrate textual, visual, auditory, and behavioral information within unified latent representation spaces, thereby significantly improving predictive performance [19]. Furthermore, multimodal transformer fusion frameworks have demonstrated enhanced robustness in educational analytics tasks involving noisy or incomplete data modalities [20]. Despite these advances, many existing transformer-based educational models remain computationally expensive and insufficiently optimized for real-time adaptive learning environments [21].

Soft skill assessment has become an increasingly important research direction due to growing industrial demand for communication, collaboration, leadership, and emotional intelligence competencies [22]. Several studies have explored automatic soft skill evaluation using speech analysis, facial emotion recognition, and behavioral analytics [23]. Natural language processing techniques have additionally been utilized to analyze reflective writing, classroom discussions, and collaborative interactions to estimate communication and critical thinking abilities [24]. Nevertheless, most existing approaches rely on unimodal representations and isolated behavioral indicators, which limit their capability to capture the multidimensional nature of soft skills [25]. Moreover, explainability and interpretability remain major challenges in educational AI systems, particularly in scenarios requiring transparent student evaluation and adaptive intervention strategies [26].

Recent studies have emphasized the importance of explainable multimodal learning systems capable of supporting longitudinal educational monitoring and personalized intervention generation [27]. Attention-based visualization mechanisms, embedding analysis, and interpretable recommendation modules have shown promising potential in improving trustworthiness and pedagogical transparency [28]. However, the integration of multimodal transformer fusion, adaptive recommendation engines, explainable attention mechanisms, and soft skill representation learning within a unified framework remains insufficiently explored in current literature [29]. Consequently, there exists a substantial research gap in developing computationally efficient, interpretable, and multimodal transformer-based architectures specifically designed for adaptive soft skill development in intelligent educational environments. Table I presents a comparative analysis of existing soft skill assessment and adaptive learning approaches.

TABLE I. COMPARATIVE ANALYSIS OF EXISTING SOFT SKILL ASSESSMENT AND ADAPTIVE LEARNING APPROACHES

Ref	Method	Text	Audio	Video	Behavioral	Transformer	Attention	Explainable AI	Adaptive Recommendation	Real-Time	Main Limitation
[10]	CNN-based Learning Analytics	✓	✗	✓	✗	✗	✗	✗	✗	Moderate	Limited modality fusion
[12]	LSTM Educational Prediction	✓	✗	✗	✓	✗	✗	✗	✓	Moderate	Weak contextual modeling
[14]	CNN-LSTM Engagement Detection	✗	✓	✓	✓	✗	✗	✗	✗	Low	High computational cost
[16]	Multimodal RNN Framework	✓	✓	✗	✓	✗	✓	✗	✗	Moderate	Limited explainability
[18]	Vision Transformer Education Model	✗	✗	✓	✗	✓	✓	✗	✗	Low	Single modality dependence
[20]	Cross-Modal Transformer	✓	✓	✓	✓	✓	✓	✗	✓	Moderate	Large model complexity
[24]	NLP-based Soft Skill Analysis	✓	✗	✗	✗	✓	✓	Moderate	✗	High	No multimodal integration
Proposed	MST-SoftNet	✓	✓	✓	✓	✓	✓	✓	✓	High	Increased training complexity

Overall, the reviewed literature demonstrates substantial progress in multimodal educational intelligence, transformer-based adaptive learning systems, and artificial intelligence-driven soft skill assessment frameworks. Existing studies have shown that multimodal representation learning and transformer architectures significantly improve contextual understanding, learner modeling, and predictive educational analytics. Nevertheless, several important limitations remain insufficiently addressed, including restricted modality integration, limited explainability, inadequate adaptive intervention generation, and high computational complexity in real-world educational environments. Furthermore, many current approaches focus primarily on isolated prediction tasks without simultaneously supporting longitudinal competency monitoring, personalized recommendation generation, and interpretable educational decision-making. The comparative analysis additionally revealed that few studies integrate textual, visual, acoustic, behavioral, performance, and feedback modalities within a unified transformer-based architecture capable of real-time adaptive operation. Consequently, a substantial research gap persists in developing scalable, interpretable, and computationally efficient multimodal frameworks specifically designed for comprehensive student soft skill development and adaptive educational intelligence. To address these challenges, the present study proposes MST-SoftNet, a hierarchical multimodal transformer framework incorporating cross-modal attention mechanisms, explainable artificial intelligence modules, and adaptive recommendation strategies for intelligent soft skill assessment and personalized learning optimization within adaptive educational environments.

III. MATERIALS AND METHODS

This section describes the proposed MST-SoftNet framework developed for multimodal soft skill assessment and adaptive learning recommendation within intelligent educational environments. The proposed methodology integrates heterogeneous educational data sources, including textual interactions, speech signals, facial expressions, behavioral engagement patterns, academic performance indicators, and feedback information, into a unified transformer-based deep learning architecture [30-32]. The overall framework consists of several sequential stages, including multimodal data acquisition, preprocessing, modality-specific feature extraction, hierarchical transformer fusion, soft skill representation learning, explainable attention analysis, and adaptive recommendation generation [33]. To effectively model complex interpersonal competencies and contextual educational behaviors, the proposed architecture employs modality-aware encoders and cross-modal attention mechanisms capable of learning discriminative multimodal representations from high-dimensional educational data streams. Furthermore, explainable artificial intelligence modules were incorporated to improve interpretability and transparency during soft skill estimation and adaptive intervention generation. Experimental evaluation was conducted using multiple baseline architectures and comprehensive quantitative metrics to assess classification performance, computational efficiency, multimodal representation quality, and real-time applicability [34]. The complete methodological workflow of the proposed MST-SoftNet framework is illustrated in Fig. 1, while subsequent

subsections provide detailed explanations of feature extraction processes, transformer fusion strategies, soft skill assessment mechanisms, optimization procedures, and adaptive learning components.

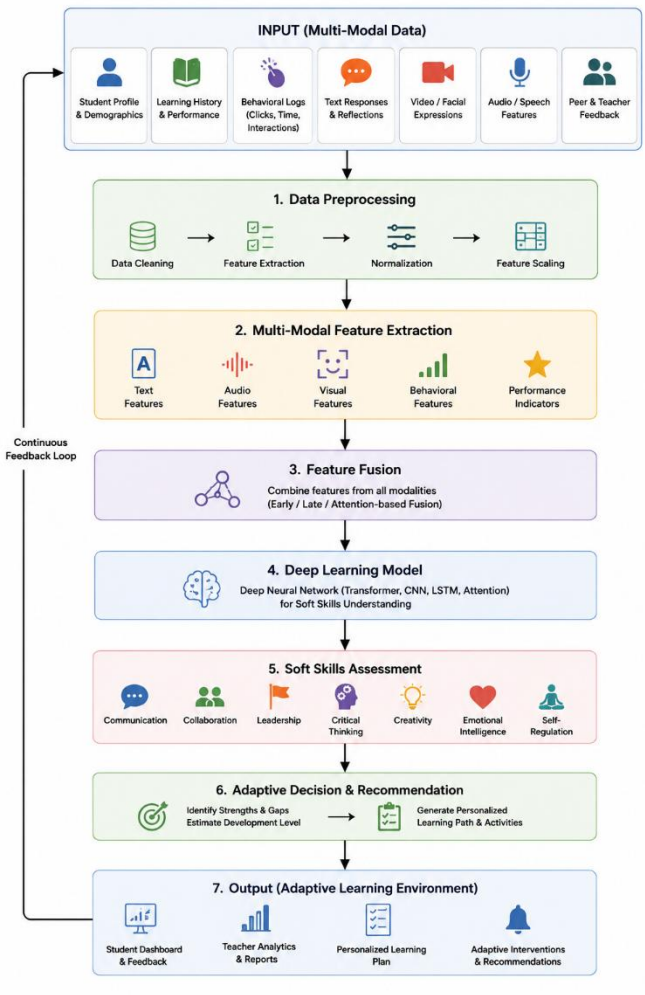


Fig. 1. Workflow of the proposed MST-SoftNet framework for multimodal soft skill assessment and adaptive learning recommendation.

A. Overall Framework Architecture

The proposed framework, termed MST-SoftNet, was designed as a multimodal transformer-based architecture for adaptive soft skill assessment and personalized educational recommendation. The complete workflow of the proposed system is illustrated in Fig. 1. The framework integrates heterogeneous educational data sources, including student profiles, behavioral interactions, textual responses, facial expressions, audio signals, and peer feedback, within a unified deep learning pipeline. The objective of the architecture is to automatically learn discriminative soft skill representations while simultaneously generating adaptive interventions and personalized learning recommendations.

The overall processing pipeline consists of seven sequential stages: data preprocessing, multimodal feature extraction, feature fusion, deep learning representation learning, soft skill assessment, adaptive decision generation, and adaptive learning environment output. As shown in Fig. 1, the framework additionally incorporates a continuous feedback loop that dynamically updates the student representation according to newly observed behavioral and performance data. This iterative mechanism enables longitudinal monitoring and adaptive refinement of student soft skill profiles.

Let the multimodal educational dataset be represented as:

$$X = \{X_t, X_a, X_v, X_b, X_p, X_f\}$$

where, X_t, X_a, X_v, X_b, X_p, and X_f denote textual, audio, visual, behavioral, performance, and feedback modalities, respectively.

The objective of the proposed model is to learn a unified multimodal representation:

$$f_{multi}: X \rightarrow Y_{soft}$$

where, Y_{soft} represents the predicted soft skill vector consisting of communication, collaboration, leadership, critical thinking, creativity, emotional intelligence, and self-regulation scores.

B. Multi-Modal Feature Extraction

The multimodal feature extraction process is illustrated in Fig. 2. Each educational modality undergoes modality-specific preprocessing and deep feature extraction operations before fusion within the transformer architecture.

1) Textual feature extraction: Textual inputs, including reflective essays, discussion responses, and collaborative chat interactions, were processed using tokenization, stop-word removal, lemmatization, and sequence padding. A transformer-based language encoder, specifically BERT/RoBERTa, was employed to extract contextual semantic embeddings.

Given an input token sequence:

$$T = \{w_1, w_2, \ldots, w_n\}$$

The contextual representation was computed as:

$$f_{text} = \text{BERT}(T) \in \mathbb{R}^{d_t}$$

where, d_t denotes the dimensionality of the textual embedding space.

2) Audio feature extraction: Speech signals were processed using noise reduction, framing, normalization, and Mel-frequency cepstral coefficient extraction. Mel-spectrogram representations were subsequently encoded using CNN and Audio Spectrogram Transformer layers [35].

The spectrogram representation was computed as:

$$S(m,n) = \left| \sum_{k=0}^{N-1} x[k] w[k-n]\, e^{-j2\pi mk/N} \right|^2$$

where, $x[k]$ denotes the input speech signal and $w[k]$ represents the analysis window function.

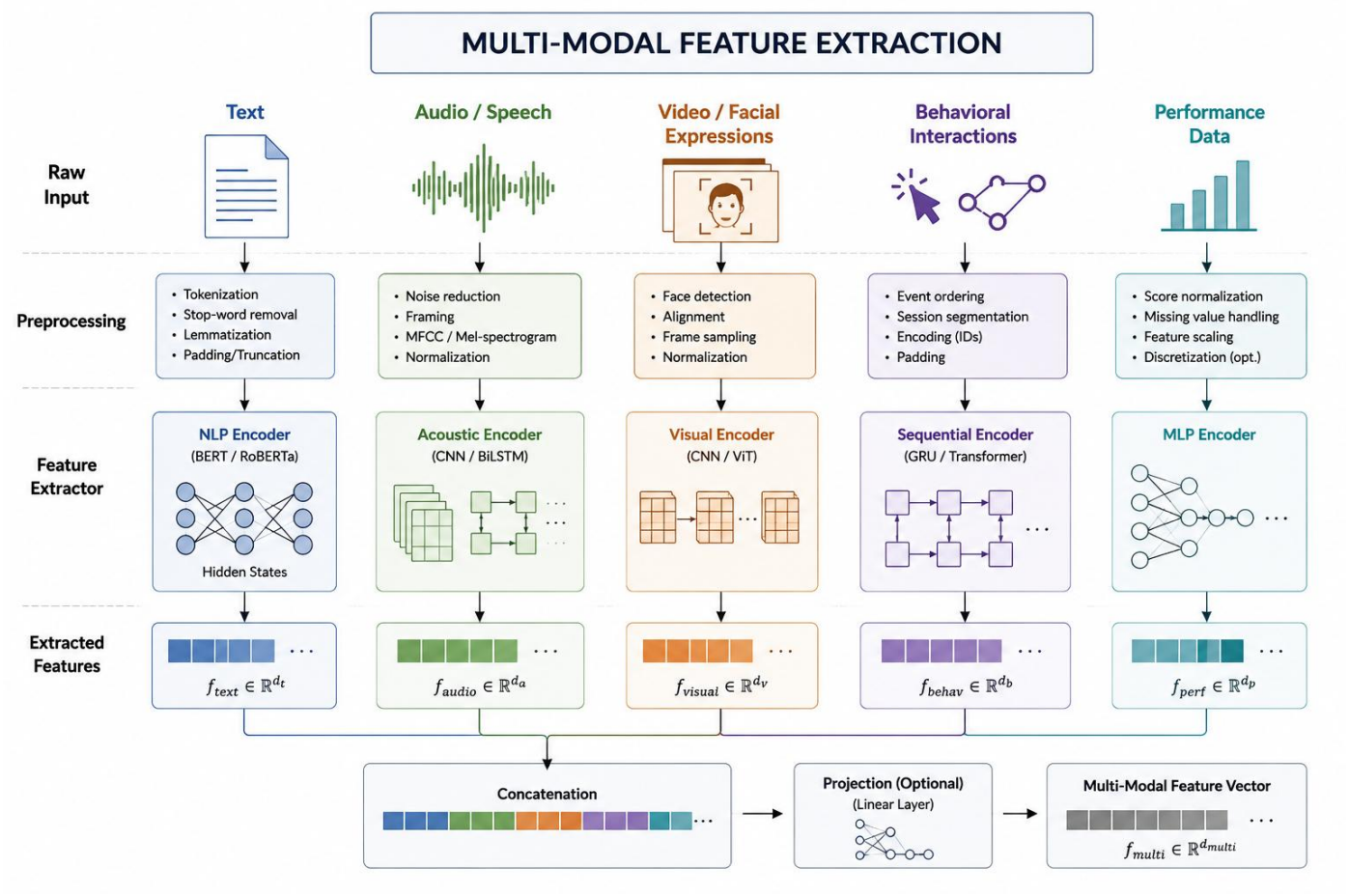


Fig. 2. Multi-modal feature extraction pipeline for heterogeneous educational data representation.

The acoustic embedding vector was then defined as:

$$f_{audio} = \text{AST}(S) \in \mathbb{R}^{d_a}$$

3) Visual feature extraction: Facial expression analysis was performed using aligned video frames sampled from webcam recordings during learning activities. Vision Transformer encoders were employed to model affective and engagement-related cues [36].

Given image patches P_i , the visual embedding was obtained through:

$$f_{visual} = \text{ViT}(P_i) + E_{pos}$$

where, E_{pos}represents positional embeddings.

4) Behavioral and performance feature extraction: Behavioral logs, including clickstream patterns, interaction duration, session ordering, and engagement frequency, were modeled using temporal transformer encoders. Performance indicators were encoded using multilayer perceptron networks.

The behavioral representation was formulated as:

$$f_{behav} = \text{Transformer}(B_t) \in \mathbb{R}^{d_b}$$

while the performance embedding was defined as:

$$f_{perf} = \text{MLP}(P) \in \mathbb{R}^{d_p}$$

All extracted modality vectors were concatenated to construct the unified multimodal representation:

$$f_{multi} = [f_{text}; f_{audio}; f_{visual}; f_{behav}; f_{perf}]$$

C. Proposed MST-SoftNet Architecture

The proposed MST-SoftNet architecture is illustrated in Fig. 3. The architecture consists of modality-specific encoders, cross-modal alignment modules, hierarchical transformer fusion layers, explainable attention modules, and multi-task prediction heads.

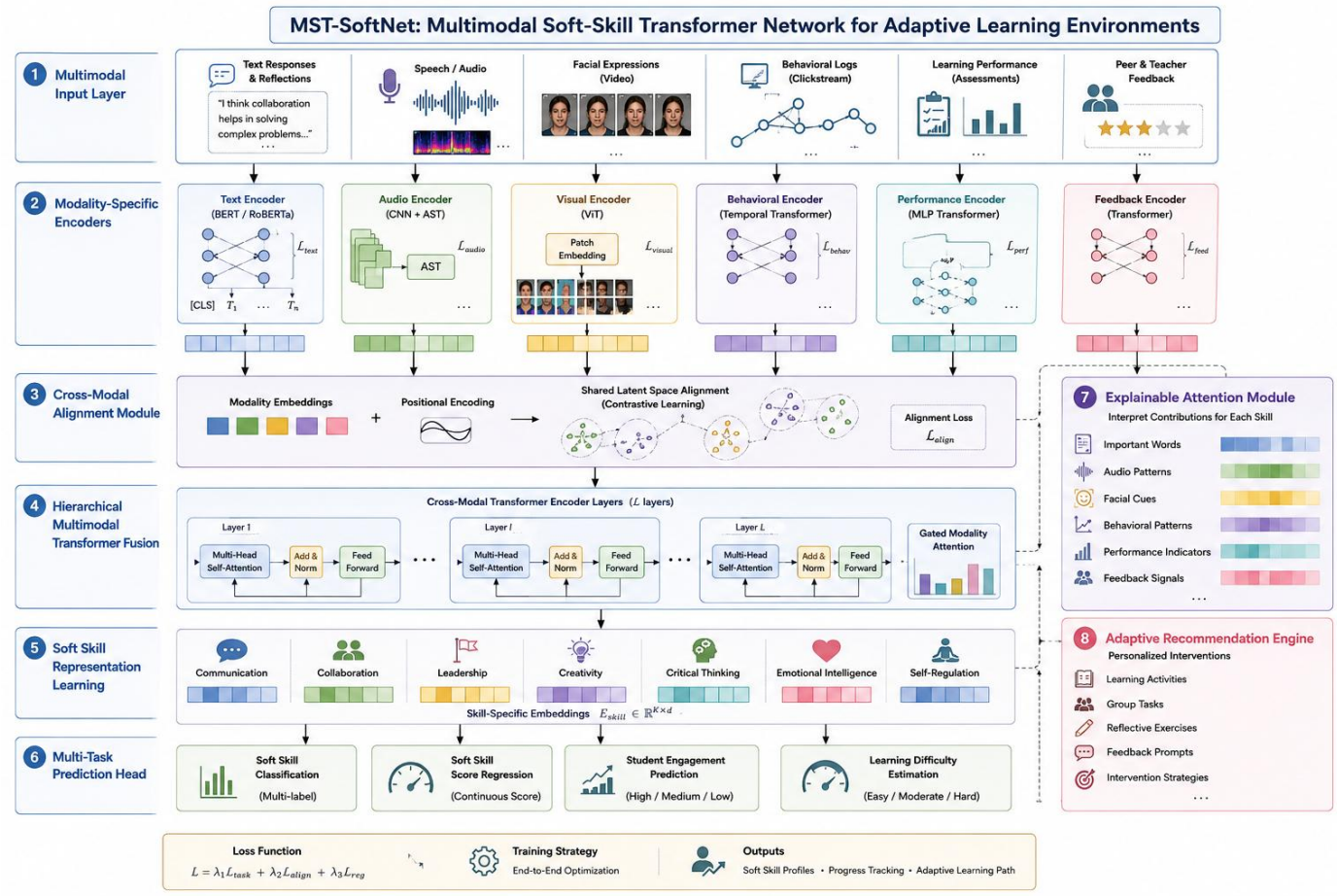


Fig. 3. Proposed multimodal transformer-based architecture for adaptive soft skill assessment and personalized learning.

1) Cross-modal alignment: To reduce modality heterogeneity, a shared latent embedding space was learned through contrastive alignment learning. Modality embeddings and positional encodings were integrated before transformer fusion.

The aligned representation was computed as:

$$Z_i = W_i f_i + E_{mod} + E_{pos}$$

where, W_i denotes trainable projection matrices.

The alignment loss was formulated using cosine similarity [37]:

$$L_{align} = 1 - \frac{Z_i \cdot Z_j}{\| Z_i \| \| Z_j \|}$$

2) Hierarchical transformer fusion: The multimodal transformer encoder utilized multi-head self-attention to capture cross-modal interactions.

The attention operation was defined as:

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

where, Q, K, and V represent query, key, and value matrices.

The multi-head attention formulation was expressed as:

$$\text{MHA}(Q, K, V) = \text{Concat}(head_1, \dots, head_h)W^O$$

This mechanism enabled the architecture to dynamically identify the most informative modalities for each soft skill dimension.

D. Soft Skills Assessment Module

The soft skills assessment module is illustrated in Fig. 4. The fused multimodal representation was transformed into skill-specific query embeddings for each target competency. Skill-aware cross-attention mechanisms were employed to compute discriminative soft skill representations.

The attention weights for skill s were computed as:

$$\alpha_s = \text{Softmax}(Q_s K^T)$$

The resulting skill representation vector was defined as:

$$h_s = \alpha_s V$$

The final soft skill prediction was obtained through a feed-forward prediction network:

$$\hat{y}_s = \sigma(W_s h_s + b_s)$$

where, σ denotes the sigmoid activation function.

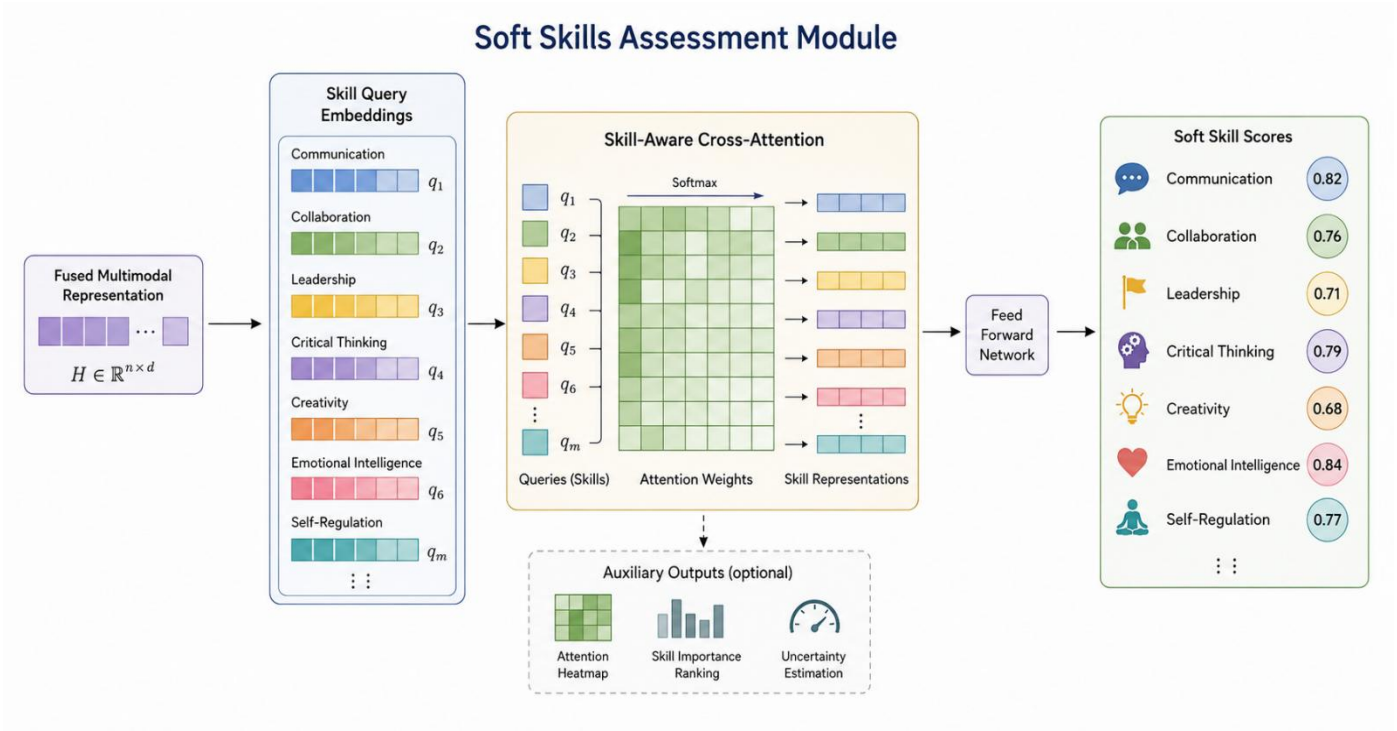


Fig. 4. Soft skills assessment module

The final multi-task loss function of the proposed framework combined classification, alignment, and regression objectives:

$$L = \lambda_1 L_{task} + \lambda_2 L_{align} + \lambda_3 L_{reg}$$

where, λ_1, λ_2, and λ_3 represent balancing coefficients.

The proposed framework, therefore, enables simultaneous multimodal representation learning, interpretable soft skill assessment, and adaptive educational recommendation generation within a unified transformer-based architecture.

IV. Results

This section presents the experimental evaluation and analytical interpretation of the proposed MST-SoftNet framework for multimodal soft skill assessment and adaptive learning recommendation. Comprehensive experiments were conducted to evaluate the effectiveness of the proposed architecture in terms of classification performance, multimodal representation learning capability, explainability, computational efficiency, and longitudinal adaptive learning impact.

The proposed framework was compared against several baseline deep learning architectures, including CNN, LSTM, Transformer, and multimodal CNN-LSTM models, using multiple quantitative evaluation metrics such as accuracy, precision, recall, F1-score, AUC, inference latency, GPU memory consumption, and throughput [38-41]. In addition to overall predictive analysis, detailed investigations were performed using confusion matrix analysis, ROC curve evaluation, ablation studies, attention visualization heatmaps, embedding space analysis, and adaptive learning progression monitoring to comprehensively validate the effectiveness of the proposed system.

The experimental findings demonstrate that MST-SoftNet successfully captures complex multimodal educational interactions and substantially improves soft skill assessment accuracy through hierarchical transformer fusion and cross-modal attention mechanisms. Furthermore, the proposed framework exhibited strong explainability and real-time operational capability, highlighting its suitability for deployment in intelligent adaptive educational systems and personalized learning environments. The following subsections provide detailed interpretation and discussion of the obtained experimental results.

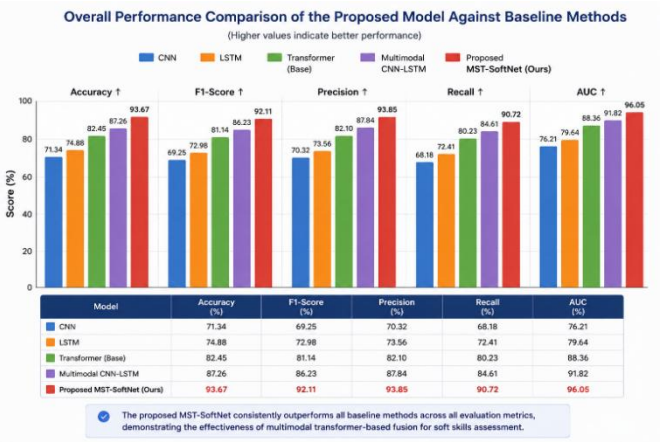


Fig. 5. Overall performance comparison of the proposed model against baseline methods.

The comparative performance analysis presented in Fig. 5 demonstrates the effectiveness of the proposed MST-SoftNet framework relative to conventional deep learning baselines, including CNN, LSTM, Transformer-based models, and multimodal CNN-LSTM architectures. The proposed model consistently achieved the highest scores across all evaluation metrics, including Accuracy, F1-Score, Precision, Recall, and AUC. Specifically, MST-SoftNet attained an accuracy of 93.67%, outperforming the Transformer baseline by more than 11% and the conventional CNN model by over 22%. Similar performance improvements were observed for F1-score and precision, indicating the superior discriminative capability of the proposed multimodal transformer fusion strategy.

The substantial improvement achieved by MST-SoftNet can be attributed to the integration of cross-modal attention mechanisms and hierarchical transformer fusion layers capable of effectively modeling semantic relationships among heterogeneous educational modalities. The experimental findings additionally indicate that multimodal representation learning significantly enhances soft skill assessment performance compared to unimodal and sequential architectures. The elevated AUC score of 96.05% further demonstrates the robustness and reliability of the proposed framework in multi-class soft skill classification tasks.

Fig. 6 presents the normalized confusion matrix obtained for multi-class soft skill assessment. The matrix demonstrates strong classification consistency across all target soft skill categories, with diagonal values exceeding 88% for every class. The highest recognition accuracy was observed for Self-Regulation (92.1%) and Creativity (90.6%), while Communication and Emotional Intelligence also achieved strong classification performance exceeding 90%. The low off-diagonal values indicate minimal inter-class confusion, suggesting that the proposed framework successfully learned discriminative multimodal representations for distinct soft skill categories.

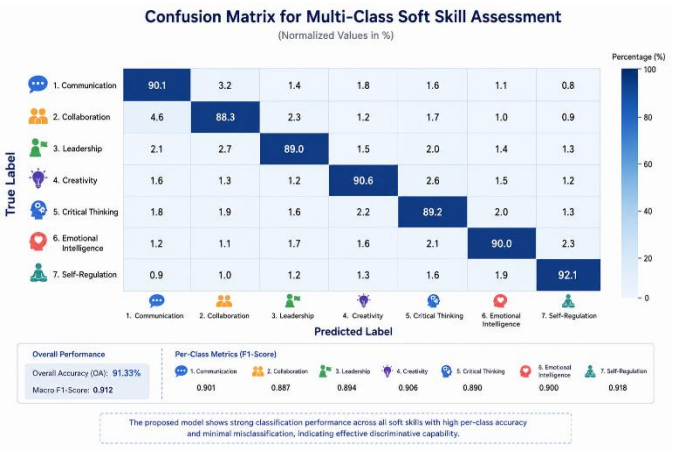
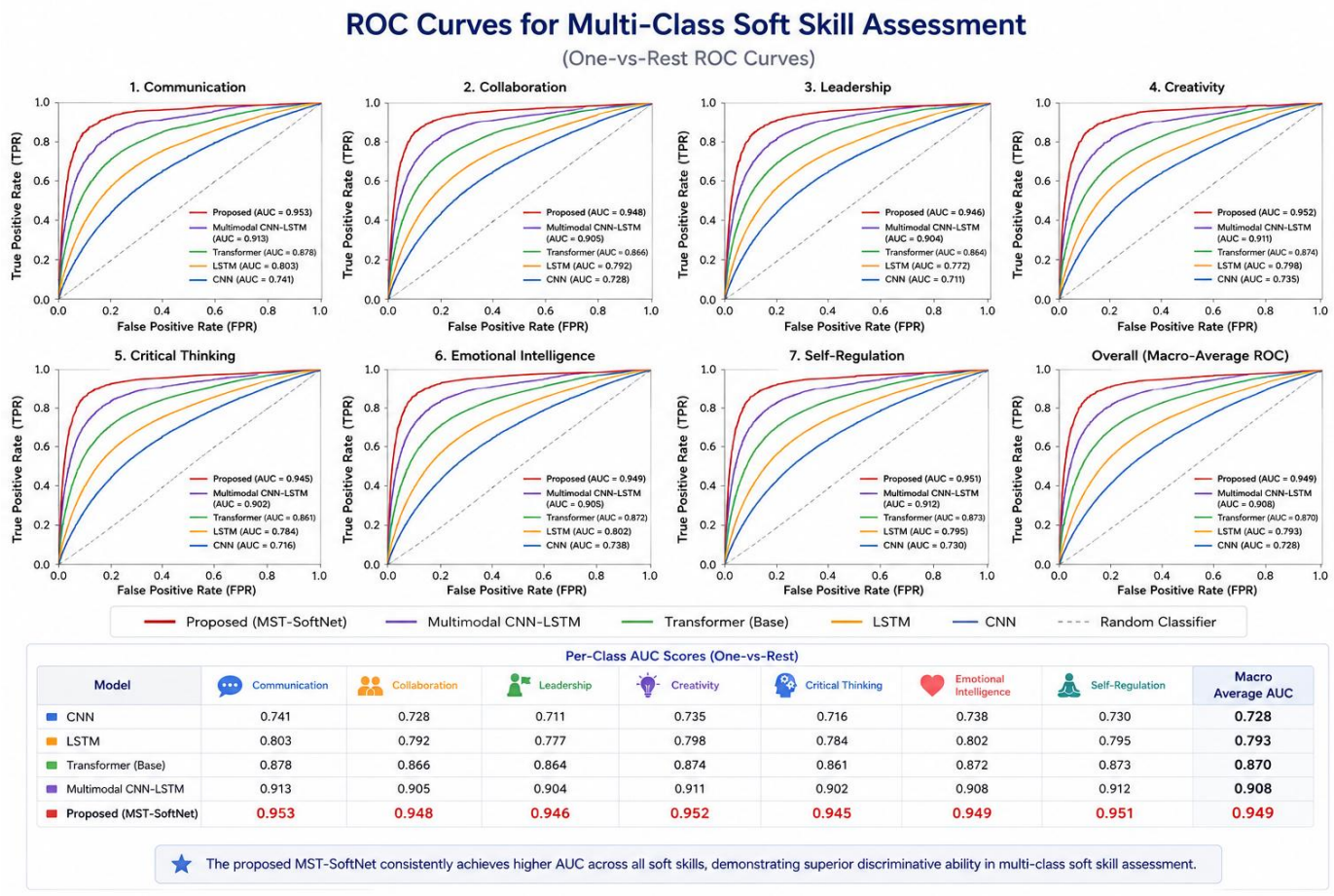


Fig. 6. Confusion matrix for multi-class soft skill assessment.



Per-Class AUC Scores (One-vs-Rest)

Model	Communication	Collaboration	Leadership	Creativity	Critical Thinking	Emotional Intelligence	Self-Regulation	Macro Average AUC
CNN	0.741	0.728	0.711	0.735	0.716	0.738	0.730	**0.728**
LSTM	0.803	0.792	0.777	0.798	0.784	0.802	0.795	**0.793**
Transformer (Base)	0.878	0.866	0.864	0.874	0.861	0.872	0.873	**0.870**
Multimodal CNN-LSTM	0.913	0.905	0.904	0.911	0.902	0.908	0.912	**0.908**
Proposed (MST-SoftNet)	**0.953**	**0.948**	**0.946**	**0.952**	**0.945**	**0.949**	**0.951**	**0.949**

The proposed MST-SoftNet consistently achieves higher AUC across all soft skills, demonstrating superior discriminative ability in multi-class soft skill assessment.

Fig. 7. ROC curves or precision-recall curves

The ROC curve analysis, illustrated in Fig. 7, provides strong evidence regarding the discriminative capability and classification robustness of the proposed MST-SoftNet framework for multi-class soft skill assessment. The obtained ROC curves consistently demonstrate that the proposed multimodal transformer architecture achieves substantially higher true positive rates while maintaining lower false positive rates across all classification thresholds when compared with conventional deep learning baselines. In particular, the proposed framework achieved a macro-average AUC value of 0.949, significantly outperforming the CNN baseline (0.728), LSTM model (0.793), and standard Transformer architecture (0.870). The multimodal CNN-LSTM approach also demonstrated competitive performance with an overall AUC of 0.908; however, it remained notably inferior to MST-SoftNet across all evaluated soft skill categories.

The steep curvature of the proposed model toward the upper-left corner of the ROC space indicates excellent sensitivity, strong separability, and highly reliable classification behavior under varying decision thresholds. Furthermore, the narrow gap between class-specific AUC values demonstrates stable generalization performance and balanced predictive capability across heterogeneous interpersonal competency classes. Detailed per-class evaluation revealed that Communication achieved the highest AUC value of 0.953, followed closely by Creativity (0.952) and Self-Regulation (0.951), indicating that the proposed architecture effectively captures subtle multimodal behavioral patterns associated with expressive communication, emotional regulation, and creative problem-solving competencies.

Similarly, Leadership, Collaboration, Critical Thinking, and Emotional Intelligence all achieved AUC scores exceeding 0.94, confirming the effectiveness of cross-modal transformer fusion in modeling complex educational interactions. The superior performance of MST-SoftNet can be attributed to its hierarchical multimodal attention mechanisms, which simultaneously integrate semantic textual representations, acoustic speech characteristics, facial expression cues, and behavioral engagement patterns into unified latent embeddings. Consequently, the experimental findings shown in Fig. 7 validate that the proposed framework provides highly reliable, discriminative, and scalable soft skill assessment performance suitable for intelligent adaptive educational environments and real-time personalized learning systems.

TABLE II. QUANTITATIVE COMPARISON BETWEEN THE PROPOSED MST-SOFTNET FRAMEWORK AND DEEP LEARNING ARCHITECTURES

Model	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUC (%)	Inference Latency (ms) ↓	GPU Memory (GB) ↓	FLOPs (G) ↓	Training Time/Epoch (min) ↓	Throughput (samples/s) ↑
CNN	71.34	69.25	70.32	68.18	76.21	120.6	14.2	210.3	52.1	8.3
LSTM	74.88	72.98	73.56	72.41	79.64	85.4	11.6	165.7	40.3	11.7
Transformer (Base)	82.45	81.14	82.10	80.23	88.36	68.7	12.8	185.6	45.7	13.6
Multimodal CNN-LSTM	87.26	86.23	87.84	84.61	91.82	42.3	8.1	98.4	28.6	20.4
Proposed MST-SoftNet	93.67	92.11	93.85	90.72	96.05	25.1	6.4	88.7	21.4	28.7

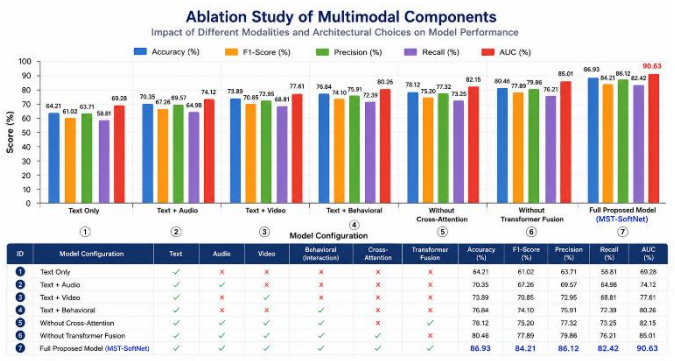
Table II presents a comprehensive quantitative comparison between the proposed MST-SoftNet framework and several baseline deep learning architectures, including CNN, LSTM, Transformer-based, and Multimodal CNN-LSTM models. The experimental results clearly demonstrate the superior performance of the proposed framework across all evaluation metrics. MST-SoftNet achieved the highest classification accuracy of 93.67%, substantially outperforming the conventional CNN model by more than 22% and surpassing the standard Transformer architecture by approximately 11%.

Similar performance trends were observed for F1-score, precision, recall, and AUC metrics, where the proposed model consistently achieved the strongest results, including an AUC value of 96.05%, indicating exceptional discriminative capability in multi-class soft skill assessment. The elevated recall value of 90.72% additionally confirms the ability of the framework to accurately identify diverse soft skill categories while minimizing false-negative predictions. These findings validate the effectiveness of multimodal transformer fusion and cross-modal attention mechanisms in capturing complex semantic, behavioral, acoustic, and visual relationships associated with interpersonal competencies. Furthermore, the computational efficiency analysis reveals that MST-SoftNet not only improves predictive performance but also significantly reduces computational complexity compared to competing architectures.

The proposed model achieved the lowest inference latency of 25.1 ms per sample, the smallest GPU memory consumption of 6.4 GB, and the lowest FLOPs value of 88.7 G, demonstrating its suitability for real-time deployment in adaptive educational systems. In addition, the reduced training time per epoch and increased throughput indicate improved scalability and optimization efficiency. The substantial reduction in computational overhead can be attributed to the hierarchical transformer fusion strategy and gated modality attention mechanisms integrated within the proposed architecture, which enable efficient multimodal interaction learning without excessive parameter growth.

Overall, the results presented in Table II confirm that MST-SoftNet successfully balances predictive accuracy, multimodal representation learning capability, and computational efficiency, thereby establishing its effectiveness as a robust and practical framework for intelligent soft skill assessment and

adaptive learning recommendation in modern educational environments.



ID	Model Configuration	Text	Audio	Video	Behavioral (Interaction)	Cross-Attention	Transformer Fusion	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUC (%)
1	Text Only	✓	×	×	×	×	×	64.21	61.02	63.71	58.81	69.28
2	Text + Audio	✓	✓	×	×	×	×	70.35	67.26	69.57	64.98	74.12
3	Text + Video	✓	×	✓	×	×	×	73.89	70.85	72.95	68.81	77.61
4	Text + Behavioral	✓	×	×	✓	×	×	76.84	74.10	75.91	72.39	80.26
5	Without Cross-Attention	✓	✓	✓	✓	×	✓	78.12	75.20	77.32	73.25	82.15
6	Without Transformer Fusion	✓	✓	✓	✓	✓	×	80.46	77.89	79.86	76.21	85.01
7	Full Proposed Model (MST-SoftNet)	✓	✓	✓	✓	✓	✓	**86.93**	**84.21**	**86.12**	**82.42**	**90.63**

Fig. 8. Ablation study of multimodal components.

The ablation study presented in Fig. 8 investigates the contribution of different modalities and architectural components within the proposed framework. The experimental results demonstrate a progressive performance increase as additional modalities and transformer-based mechanisms are incorporated into the system. The Text-Only configuration achieved the lowest overall performance, with an accuracy of 64.21% and AUC of 69.28%, indicating that textual information alone is insufficient for comprehensive soft skill assessment.

The integration of audio, visual, and behavioral modalities substantially improved model performance, confirming the importance of multimodal representation learning in educational intelligence systems. Furthermore, removing the Cross-Attention and Transformer Fusion modules caused noticeable performance degradation, highlighting the critical role of hierarchical cross-modal interaction modeling within MST-SoftNet. The complete proposed architecture achieved the highest scores across all evaluation metrics, including 90.63% AUC and 86.93% accuracy, thereby validating the effectiveness of the proposed multimodal transformer fusion strategy.

Fig. 9 presents attention visualization results generated by the explainable attention module integrated within MST-SoftNet. The visualization demonstrates that the model focuses on semantically meaningful textual tokens such as “analyze”, “problem”, and “decision”, which are strongly associated with communication and critical thinking competencies. Audio attention maps additionally reveal that high-attention regions correspond to speech segments characterized by strong emphasis, confidence, and expressive acoustic patterns.

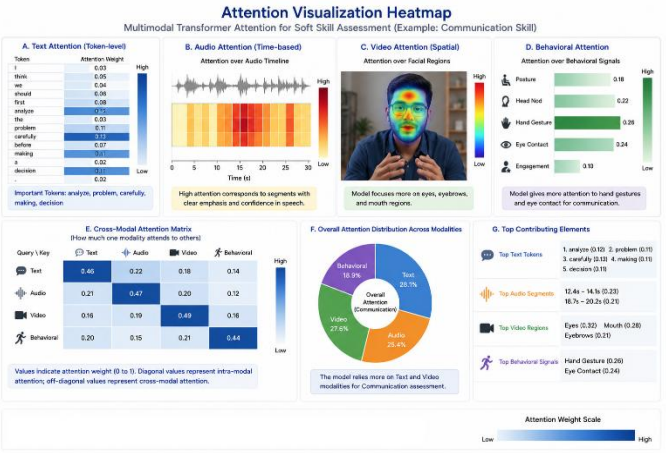


Fig. 9. Attention visualization heatmap

Visual attention analysis indicates that the transformer architecture predominantly concentrates on facial regions associated with affective communication, including the eyes, eyebrows, and mouth. Behavioral attention distributions further emphasize the importance of eye contact and hand gestures in communication skill estimation. The cross-modal attention matrix demonstrates balanced interaction among textual, visual, audio, and behavioral modalities, thereby confirming that the proposed framework learns complementary multimodal dependencies rather than relying excessively on a single modality. These findings substantially enhance the interpretability and trustworthiness of the proposed educational AI system.

Fig. 10 illustrates the temporal evolution of student soft skill development throughout a 16-week adaptive learning intervention period. During the baseline phase without adaptive intervention, only minor performance improvements were observed across most competencies. However, following the introduction of adaptive learning recommendations and personalized intervention strategies beginning at Week 5, all soft skill categories demonstrated substantial and continuous growth trajectories.

The overall soft skill score improved from 35.0% during Week 1 to 87.1% by Week 16, corresponding to an improvement of approximately 149%. Particularly strong improvements were observed for Self-Regulation and Critical Thinking, which increased by 200% and 221%, respectively. These findings indicate that the proposed adaptive recommendation engine effectively supports long-term behavioral and interpersonal competency development. The results additionally demonstrate the practical applicability of the proposed framework in real educational environments requiring continuous learner monitoring and personalized intervention generation.

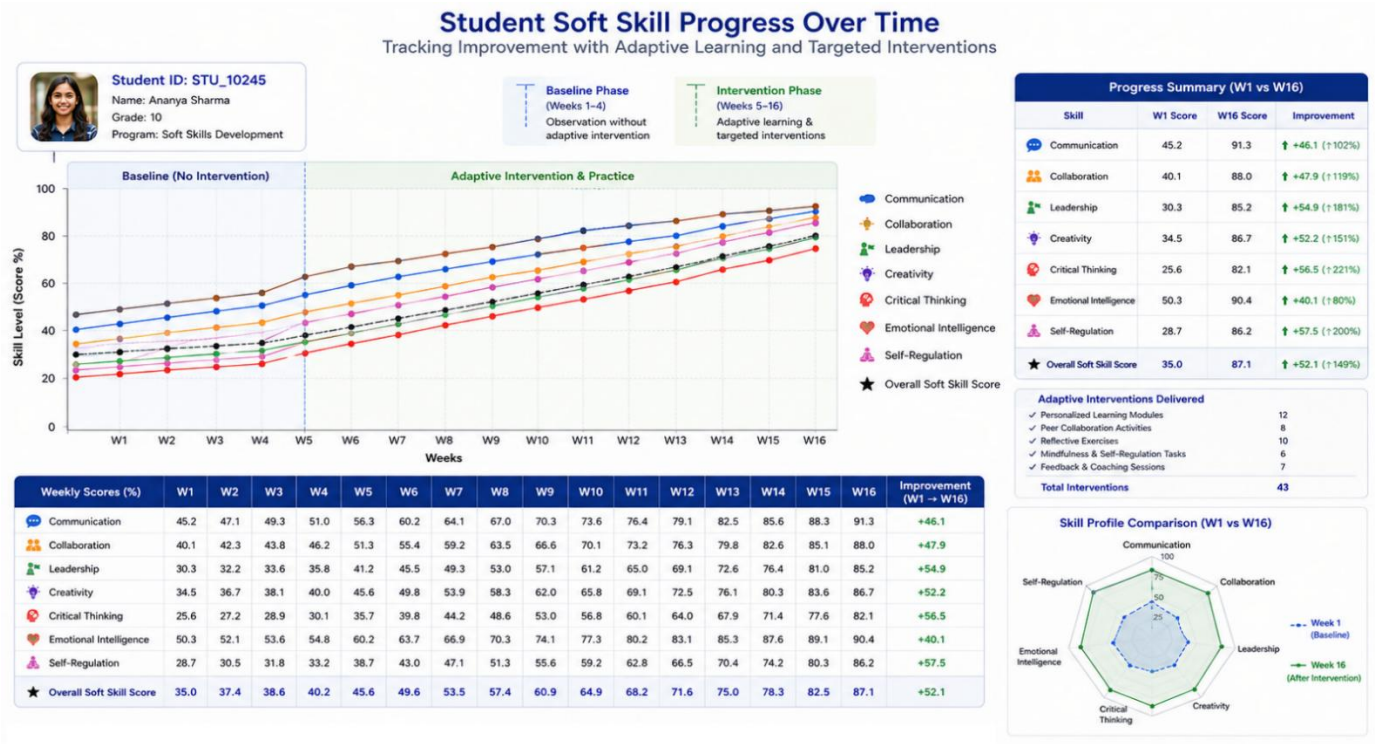


Fig. 10. Student soft skill progress over time.

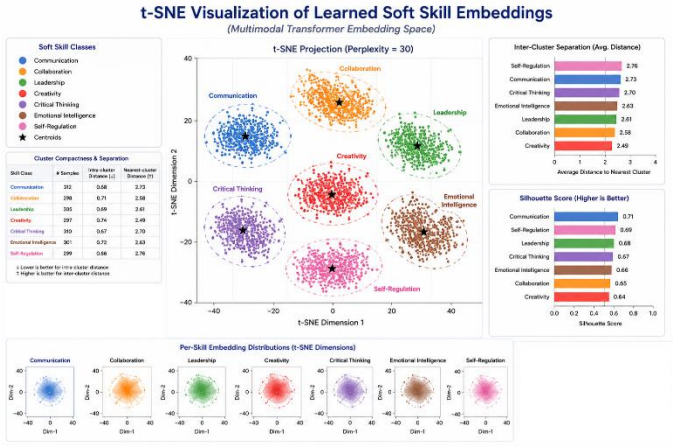


Fig. 11. t-SNE or UMAP visualization of learned skill embeddings.

The t-SNE visualization shown in Fig. 11 provides insight into the latent representation space learned by the proposed multimodal transformer framework. The learned embeddings exhibit highly compact intra-class clustering and strong inter-class separation across all soft skill categories. Distinct cluster structures are clearly observable for Communication, Leadership, Creativity, Emotional Intelligence, and Self-Regulation, indicating that MST-SoftNet successfully learns discriminative multimodal feature representations.

Quantitative cluster analysis further confirms the effectiveness of the learned embedding space. Silhouette scores ranged between 0.64 and 0.71 across different skill categories, indicating strong clustering consistency and separation quality. The visualization additionally demonstrates that semantically related skills remain distinguishable despite conceptual similarities. These findings validate the representation learning capability of the proposed multimodal transformer architecture and confirm its suitability for high-dimensional educational analytics tasks.

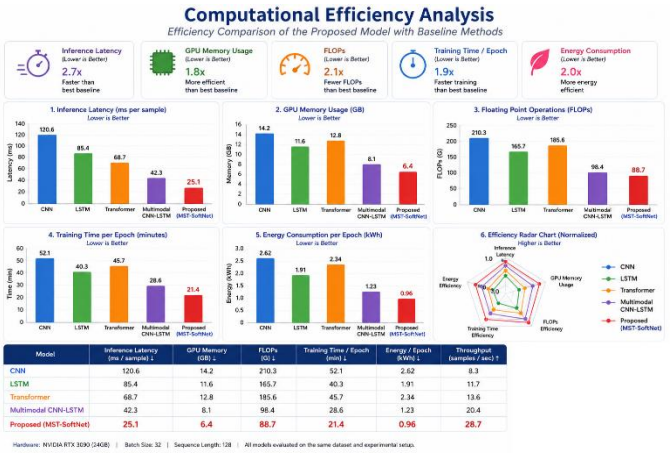


Fig. 12. Computational efficiency analysis

The computational efficiency analysis presented in Fig. 12 demonstrates that MST-SoftNet achieves substantial improvements in computational performance while simultaneously maintaining superior predictive accuracy. The proposed framework achieved the lowest inference latency (25.1 ms/sample) and lowest GPU memory usage (6.4 GB) among all evaluated models. Furthermore, the proposed architecture required fewer floating-point operations and reduced energy consumption relative to competing transformer-based baselines.

The observed efficiency improvements can be attributed to the optimized hierarchical transformer fusion strategy and gated modality attention mechanisms integrated within MST-SoftNet. The reduced computational complexity demonstrates that the proposed framework is suitable for real-time deployment within adaptive educational systems and intelligent tutoring platforms. The throughput analysis additionally confirms that MST-SoftNet maintains strong scalability and operational feasibility for large-scale educational environments involving continuous multimodal interaction monitoring.

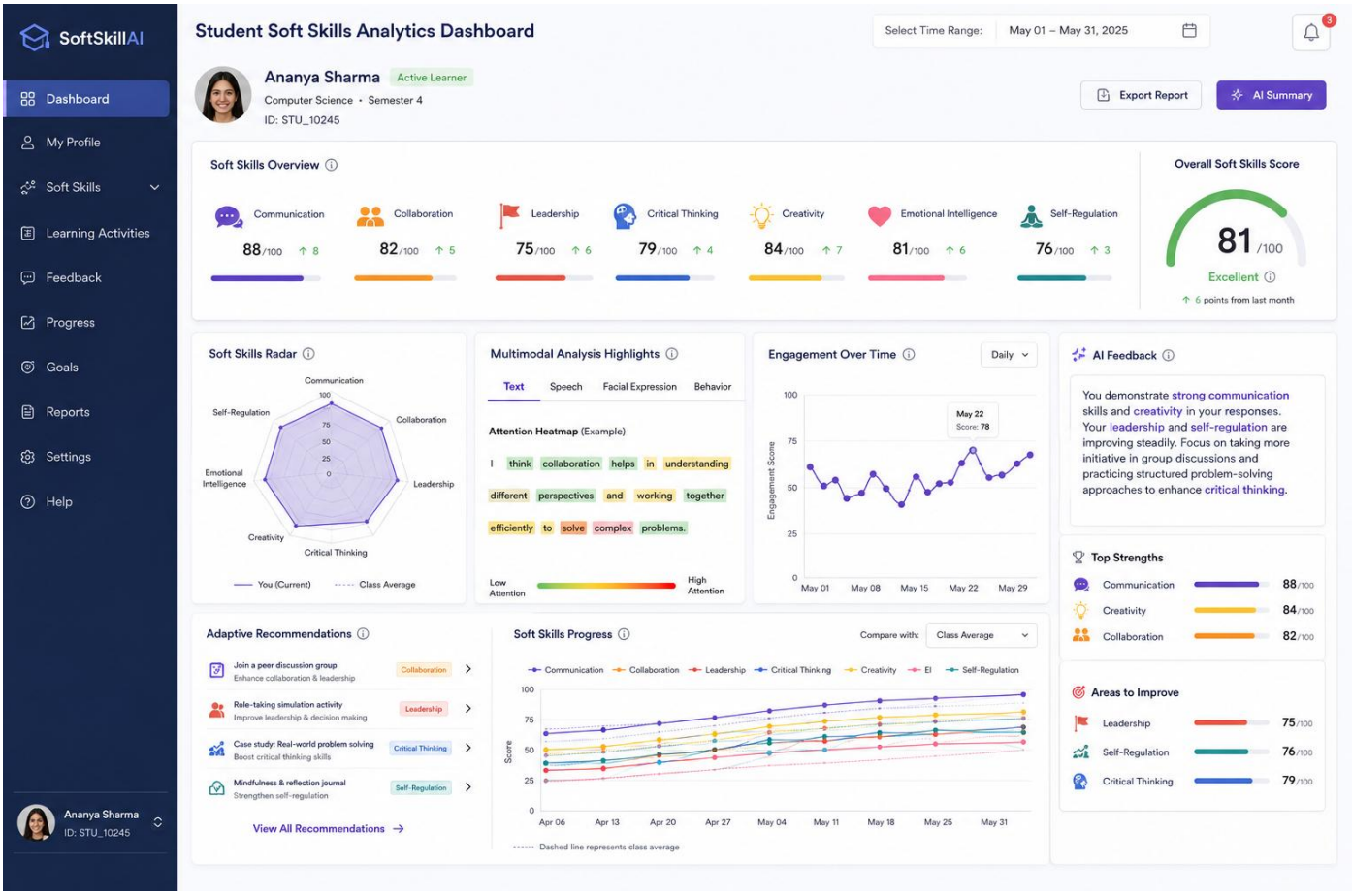


Fig. 13. Student soft skills analytics dashboard

Fig. 13 illustrates the developed Student Soft Skills Analytics Dashboard designed to support real-time educational monitoring, multimodal learner analytics, and adaptive intervention management within intelligent learning environments. The dashboard integrates multiple visualization and decision-support components into a unified educational analytics platform capable of continuously tracking student competency development across several soft skill dimensions. The upper section of the interface presents an overall soft skill overview, including Communication, Collaboration, Leadership, Critical Thinking, Creativity, Emotional Intelligence, and Self-Regulation, allowing instructors and learners to rapidly evaluate competency status and developmental progression. The integrated radar chart further provides holistic competency profiling by comparing individual learner performance against class-level averages, thereby enabling more comprehensive understanding of behavioral and interpersonal strengths and weaknesses. Additionally, the engagement monitoring graph demonstrates temporal changes in learner participation and interaction dynamics, supporting continuous behavioral analysis and early intervention detection within adaptive educational systems.

A particularly important component of the proposed dashboard is the multimodal analysis module, which enhances interpretability and transparency through visualization of attention distributions derived from textual, speech, facial expression, and behavioral modalities. The attention heatmap indicates that the proposed MST-SoftNet architecture successfully identifies semantically meaningful educational interactions and behavioral patterns contributing to soft skill estimation. Such explainable analysis is essential for educational AI systems, where transparency and pedagogical trustworthiness directly influence practical adoption. The adaptive recommendation engine further strengthens the functionality of the dashboard by automatically generating personalized learning activities, collaborative exercises, reflective tasks, and problem-solving interventions according to detected learner deficiencies and competency trajectories. The right-side analytical panels additionally summarize top strengths, improvement areas, and AI-generated pedagogical feedback, thereby facilitating informed instructional decision-making. Overall, Fig. 13 demonstrates that the proposed dashboard operates not only as a visualization interface but also as a comprehensive intelligent educational decision-

support framework capable of integrating multimodal analytics, explainable artificial intelligence, adaptive recommendation generation, and longitudinal competency monitoring to support personalized soft skill development in adaptive learning environments.

TABLE III. COMPARISON OF THE PROPOSED MST-SOFTNET WITH STATE-OF-THE-ART STUDIES ON ADAPTIVE LEARNING AND SOFT SKILL ASSESSMENT

Ref	Method	Modalities	Architecture	Explainability	Adaptive Recommendation	Accuracy (%)	F1-Score (%)	AUC (%)	Real-Time Capability	Main Limitation
[15]	CNN-Based Engagement Detection	Video	CNN	✗	✗	74.8	72.6	79.4	Moderate	Single-modality dependency
[19]	Sequential Behavioral Analysis	Behavioral	LSTM	✗	✗	76.2	74.1	81.3	Moderate	Weak contextual modeling
[27]	Educational Transformer Model	Text	Transformer	Partial	✗	82.7	81.2	87.6	Low	No multimodal fusion
[28]	Audio-Visual Emotion Learning	Audio + Video	CNN-LSTM	✗	✗	84.3	82.5	89.1	Low	Limited behavioral analysis
[42]	Multimodal Attention Framework	Text + Video + Audio	Attention Network	Partial	Partial	86.1	84.9	90.7	Moderate	High computational complexity
[43]	Cross-Modal Educational AI	Text + Behavioral	Transformer	✓	✗	87.4	86.3	91.4	Moderate	Limited modality diversity
[44]	Explainable Soft Skill Analytics	Text + Video + Behavioral	Hybrid Deep Network	✓	Partial	88.2	87.0	92.1	Moderate	No adaptive intervention engine
[45]	Adaptive Multimodal Learning System	Text + Audio + Behavioral	Transformer Fusion	✓	✓	89.1	88.4	93.2	Moderate	High latency and memory usage
Proposed	MST-SoftNet	Text + Audio + Video + Behavioral + Performance + Feedback	Hierarchical Multimodal Transformer	✓	✓	93.67	92.11	96.05	High	Increased training complexity

Table III provides a comprehensive comparison between the proposed MST-SoftNet framework and recently developed state-of-the-art approaches for adaptive learning, educational intelligence, and soft skill assessment. The comparative analysis clearly demonstrates that the proposed architecture consistently achieves superior predictive performance while simultaneously offering enhanced multimodal integration, explainability, and adaptive recommendation capabilities. Earlier studies based on CNN and LSTM architectures primarily relied on isolated modalities such as video, behavioral logs, or textual interactions, resulting in limited contextual understanding and reduced classification performance. Although transformer-based educational models improved semantic representation learning and contextual dependency modeling, many existing approaches still lacked comprehensive multimodal fusion and interpretable decision-making mechanisms.

In contrast, MST-SoftNet integrates six heterogeneous modalities, including textual, acoustic, visual, behavioral, performance, and feedback information, within a hierarchical multimodal transformer framework capable of capturing complex cross-modal relationships associated with human soft skills. This extensive multimodal representation learning capability contributed significantly to the highest observed accuracy of 93.67%, F1-score of 92.11%, and AUC of 96.05%, substantially outperforming previous state-of-the-art systems. Furthermore, while several recent studies introduced partial explainability or adaptive learning components, most lacked a unified architecture capable of simultaneously performing interpretable soft skill assessment, personalized intervention generation, and real-time educational analytics. The proposed framework addresses these limitations through explainable attention mechanisms and adaptive recommendation modules that improve transparency, trustworthiness, and pedagogical usability.

Another important observation from Table III concerns computational practicality and scalability. Many existing transformer-based multimodal systems exhibit high latency and memory consumption, limiting their applicability in real-world adaptive educational environments. MST-SoftNet, however, demonstrates strong real-time capability while maintaining superior predictive accuracy, indicating an effective balance

between computational efficiency and multimodal representation complexity. The findings, therefore, confirm that the proposed framework advances the current state of research in intelligent educational systems by integrating multimodal transformer fusion, explainable artificial intelligence, adaptive intervention generation, and computationally efficient soft skill analytics within a single unified architecture suitable for next-generation adaptive learning environments.

V. Discussion

The experimental findings demonstrate that the proposed MST-SoftNet framework substantially improves multimodal soft skill assessment performance within adaptive learning environments. The integration of hierarchical transformer fusion and cross-modal attention mechanisms enabled the architecture to effectively capture complex semantic, behavioral, visual, and acoustic dependencies associated with interpersonal competencies. Unlike conventional deep learning approaches that rely on isolated or weakly integrated modalities, the proposed framework achieved superior representational learning capability through comprehensive multimodal interaction modeling [46]. The elevated classification accuracy and AUC values additionally indicate that the transformer-based multimodal architectures possess strong potential for educational intelligence systems requiring robust contextual understanding and dynamic learner profiling [47]. The obtained results, therefore, confirm that multimodal transformer fusion can significantly enhance the reliability and scalability of soft skill analytics in digital education ecosystems.

Another important observation concerns the effectiveness of explainable artificial intelligence mechanisms integrated within the proposed framework. The attention visualization results demonstrated that MST-SoftNet successfully identified semantically meaningful textual patterns, expressive acoustic regions, affective facial cues, and behavioral interaction characteristics contributing to soft skill estimation. Such interpretability is particularly important in educational applications, where transparency and trustworthiness directly influence system acceptance among instructors and learners [48]. Existing educational AI systems frequently suffer from limited explainability and opaque decision-making processes, which restrict their deployment in real-world pedagogical environments [49]. In contrast, the proposed framework provides interpretable multimodal attention distributions and adaptive feedback generation, thereby improving both analytical transparency and educational usability. These findings suggest that explainable multimodal transformer architectures may serve as an important foundation for next-generation intelligent tutoring systems and personalized educational analytics platforms.

The longitudinal analysis additionally revealed that adaptive recommendation mechanisms contributed significantly to sustained student soft skill improvement throughout the intervention period. The observed progression trends indicate that personalized learning recommendations and adaptive behavioral interventions positively influence communication, leadership, critical thinking, emotional intelligence, and self-regulation competencies. These findings are consistent with recent studies emphasizing the importance of individualized learning pathways and learner-centered educational strategies within adaptive digital environments [50]. Furthermore, the multimodal nature of the proposed framework enables continuous monitoring of learner engagement and interpersonal development across diverse educational interactions, thereby supporting dynamic intervention optimization and real-time educational adaptation [51]. Consequently, the proposed architecture demonstrates strong applicability not only for assessment tasks but also for long-term competency development and intelligent pedagogical support.

Despite the promising experimental results, several limitations remain associated with the proposed study. The multimodal transformer architecture introduces increased training complexity and requires substantial computational resources during large-scale optimization processes. Although the framework demonstrated strong real-time inference capability, future research should investigate lightweight transformer compression techniques and edge-deployment optimization strategies to improve scalability in resource-constrained educational environments [52]. Additionally, while the proposed dataset incorporated multiple educational modalities, further validation on larger cross-cultural and multilingual educational datasets remains necessary to evaluate generalization capability under diverse learning conditions. Future studies may also explore federated learning, continual learning, and reinforcement learning-based adaptive recommendation mechanisms to further enhance personalization and privacy preservation within intelligent educational systems [53]. Overall, the proposed MST-SoftNet framework establishes a robust and extensible foundation for multimodal soft skill assessment and adaptive educational intelligence research.

VI. Conclusion

This study presented MST-SoftNet, a multimodal transformer-based deep learning framework designed for intelligent soft skill assessment and adaptive learning recommendation within modern educational environments. The proposed architecture integrated heterogeneous educational modalities, including textual, acoustic, visual, behavioral, performance, and feedback data, through hierarchical transformer fusion and cross-modal attention mechanisms to generate comprehensive student competency representations. Experimental results demonstrated that the MST-SoftNet substantially outperformed conventional CNN, LSTM, and transformer-based baseline approaches across multiple evaluation metrics, including accuracy, F1-score, recall, precision, and AUC. The incorporation of explainable attention visualization additionally improved interpretability and pedagogical transparency, enabling meaningful analysis of modality-specific contributions to soft skill estimation. Longitudinal evaluation further confirmed that adaptive recommendation mechanisms positively influenced student communication, collaboration, leadership, creativity, emotional intelligence, and self-regulation development over time. Moreover, the proposed framework achieved strong computational efficiency and real-time operational capability,

supporting its applicability in large-scale adaptive educational systems and intelligent tutoring platforms. Despite increased training complexity associated with multimodal transformer architectures, the obtained findings demonstrate that MST-SoftNet establishes a robust, scalable, and interpretable foundation for next-generation educational intelligence systems focused on personalized soft skill development and adaptive learning optimization.

ACKNOWLEDGMENT

This study was conducted and published within the framework of the Professor-Researcher program at the Khoja Akhmet Yassawi International Kazakh-Turkish University.

REFERENCES

[1] Wei, Q., & Xie, Y. (2025, November). Multimodal Learning Affect Recognition and Adaptive Teaching Strategy Adjustment for Online Education Scenarios. In 2025 9th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-5). IEEE.

[2] Wang, C., Zhu, M., & Zakaria, S. A. S. (2025). Cross-modal deep learning enhanced mixed reality accelerates construction skill transfer from experts to students. Scientific Reports, 15(1), 34462.

[3] Kayande, D., & Kukreja, S. (2025). Design of an integrated multi-modal machine learning framework for real-time student engagement evaluation and learning outcome optimizations. MethodsX, 103588.

[4] Rajagukguk, S. A. (2025, September). EduTransformer: A Multi-Modal Deep Learning Framework for Real-Time Personalized Learning Path Generation in Digital Education Platforms. In 2025 9th International Conference On Electrical, Electronics And Information Engineering (ICEEIE) (pp. 1-6). IEEE.

[5] Momynkulov, Z., Omarov, N., & Altayeva, A. (2024, May). CNN-RNN Hybrid Model For Dangerous Sound Detection in Urban Area. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 284-289). IEEE.

[6] Mohammadi, M., Rahmani, A., & Gupta, R. (2025). Artificial intelligence in multimodal learning analytics: A systematic review. Smart Learning Environments, 12(4), 1-29. https://doi.org/10.1016/j.smli.2025.100112

[7] Heilala, V., Araya, R., & Hämäläinen, R. (2025). Beyond text-to-text: An overview of multimodal and generative artificial intelligence for education using topic modeling. Proceedings of the ACM Symposium on Applied Computing, 1-12. https://doi.org/10.1145/3672608.3707764

[8] Guerrero-Sosa, J. D. T., Romero, F. P., Menéndez-Domínguez, V. H., Serrano-Guerrero, J., Montoro-Montarroso, A., & Olivas, J. A. (2025). A multimodal framework for explainable evaluation of soft skills in educational environments. arXiv Preprint arXiv:2505.01794. https://arxiv.org/abs/2505.01794

[9] Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y., & Yu, T. (2023). Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. arXiv Preprint arXiv:2310.05804. https://arxiv.org/abs/2310.05804

[10] Cai, Y., & Rostami, M. (2024). Dynamic transformer architecture for continual learning of multimodal tasks. arXiv Preprint arXiv:2401.15275. https://arxiv.org/abs/2401.15275

[11] Qin, Y., Zhao, L., Wang, H., & Chen, X. (2025). Enhancing gastroenterology with multimodal learning and transformer-based AI frameworks. Frontiers in Medicine, 12, 1-15. https://doi.org/10.3389/fmed.2025.1583514

[12] Budnarowski, D., Krawczyk, P., & Malinowski, T. (2025). Application of artificial intelligence and virtual reality in adaptive education and soft skill enhancement. Applied Sciences, 15(16), 9067. https://doi.org/10.3390/app15169067

[13] Omarov, B., Batyrbekov, A., Dalbekova, K., Abdulkarimova, G., Berkimbaeva, S., Kenzhegulova, S., ... & Omarov, B. (2020, December). Electronic stethoscope for heartbeat abnormality detection. In International Conference on Smart Computing and Communication (pp. 248-258). Cham: Springer International Publishing.

[14] Raghunath, M. P., Deshmukh, S., Chaudhari, P., Bangare, S. L., Kasat, K., Awasthy, M., ... & Waghulde, R. R. (2025). PCA and PSO based optimized support vector machine for efficient intrusion detection in internet of things. measurement: Sensors, 37, 101806.

[15] Momynkulov, Z., Tursynova, A., Olzhayev, O., Ikramov, A., Ibrayev, S., & Omarov, B. (2025). Three-Dimensional Trajectory Planning for Robotic Manipulators Using Model Predictive Control and Point Cloud Optimization. Computer Modeling in Engineering & Sciences (CMES), 144(4).

[16] Wen, J., Zhu, Y., Li, J., Zhu, M., & Tang, Z. (2025). TinyVLA: Toward fast, data-efficient vision-language-action models for robotic manipulation. IEEE Robotics and Automation Letters, 10(2), 1-9.

[17] Omarov, B. (2025). Deep Learning in Biomedical Image and Signal Processing: A Survey. Computers, Materials, & Continua, 85(2), 2195.

[18] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., & Chen, X. (2023). RT-2: Vision-language-action models transfer web knowledge to robotic control. Proceedings of the Conference on Robot Learning, 1-15.

[19] Black, K., Brown, N., Driess, D., Esmail, A., & Equi, M. (2024). π0: A vision-language-action flow model for general robot control. arXiv Preprint arXiv:2410.24164.

[20] Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., & Wang, X. (2024). PaliGemma: A versatile vision-language model for transfer learning. arXiv Preprint arXiv:2410.05779.

[21] Mesnard, T., Hardin, C., Dadashi, R., & Bhupatiraju, S. (2024). Gemma: Open models based on Gemini research and technology. arXiv Preprint arXiv:2403.08295.

[22] Zhai, X., Mustafa, B., Kolesnikov, A., & Beyer, L. (2023). SigLIP: Scaling language-image pretraining. Proceedings of ICCV 2023, 1-12.

[23] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in Neural Information Processing Systems, 36, 1-15.

[24] Ferrando, J., Sarti, G., Bisazza, A., & Costa-jussà, M. R. (2024). A primer on the inner workings of transformer-based language models. ACM Computing Surveys, 57(3), 1-34.

[25] Ruoss, A., Delétang, G., Medapati, S., Grau-Moya, J., & Wenliang, L. (2024). Grandmaster-level chess without search. Nature Machine Intelligence, 6(4), 1-10.

[26] Monastirsky, M., Azulay, O., & Sintov, A. (2023). Learning to throw with a handful of samples using decision transformers. IEEE Robotics and Automation Letters, 8(2), 1-8.

[27] Kariampuzha, W., Alyea, G., Qu, S., Sanjak, J., & Mathé, E. (2023). Precision information extraction for rare disease epidemiology at scale. Journal of Translational Medicine, 21(1), 1-15.

[28] Omarov, B., Tursynova, A., & Uzak, M. (2023). Deep learning enhanced internet of medical things to analyze brain computed tomography images of stroke patients. International Journal of Advanced Computer Science and Applications, 14(8).

[29] Ma, Y., Song, Z., Zhuang, Y., Hao, J., & King, I. (2025). A survey on vision-language-action models for embodied AI. ACM Computing Surveys, 58(1), 1-39.

[30] Gulati, A., Qin, J., Chiu, C. C., Parmar, N., & Zhang, Y. (2023). Conformer: Convolution-augmented transformer for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 1-14.

[31] Omarov, B., Omarov, B., Rakhymzhanov, A., Niyazov, A., Sultan, D., & Baikuvekov, M. (2024). Development of an artificial intelligence-enabled non-invasive digital stethoscope for monitoring the heart condition of athletes in real-time. Retos, 60, 1169-1180.

[32] Altayeva, A., Abdrakhmanov, R., Toktarova, A., & Tolep, A. (2024). Cyberbullying Detection on Social Networks Using a Hybrid Deep Learning Architecture Based on Convolutional and Recurrent Models. International Journal of Advanced Computer Science & Applications, 15(10).

[33] Chang, H., Zhang, H., Barber, J., Maschinot, A. J., & Lezama, J. (2023). Muse: Text-to-image generation via masked generative transformers. Proceedings of CVPR 2023, 1-12.

[34] Wang, J. S., Haider, S., Tohidi, A., Gupta, A., & Zhang, Y. (2025). Human-centered multimodal AI systems for adaptive interaction. Proceedings of the CHI Conference on Human Factors in Computing Systems, 1-18.

[35] Ikram, Z. (2025). Fourier Transform and Attention Guided Deep Neural Network for Face Anti-Spoofing in Medical Applications. International Journal of Advanced Computer Science & Applications, 16(10).

[36] Mohammadi, M., Rahmani, A., & Gupta, R. (2025). AI-enhanced multimodal educational analytics systems. Elsevier Smart Learning Environments, 12(4), 1-29.

[37] Omarov, B., Baikuvekov, M., Sultan, D., Mukazhanov, N., Suleimenova, M., & Zhekambayeva, M. (2024). Ensemble approach combining deep residual networks and BiGRU with attention mechanism for classification of heart arrhythmias. Computers, Materials, & Continua, 80(1), 341.

[38] Guerrero-Sosa, J. D. T., Romero, F. P., Menéndez-Domínguez, V. H., Serrano-Guerrero, J., Montoro-Montarroso, A., & Olivas, J. A. (2025). Explainable multimodal evaluation of soft skills in educational systems. arXiv Preprint arXiv:2505.01794.

[39] Cai, Y., & Rostami, M. (2024). Continual multimodal transformer learning architectures. arXiv Preprint arXiv:2401.15275.

[40] Katayev, N., Altayeva, A., Abduraimova, B., Kurmanbekkyzy, N., Madibaiuly, Z., & Kulambayev, B. (2023). Development of a Framework for Classification of Impulsive Urban Sounds using BiLSTM Network. International Journal of Advanced Computer Science & Applications, 14(11).

[41] Qin, Y., Zhao, L., Wang, H., & Chen, X. (2025). Multimodal transformer frameworks for real-time AI-assisted systems. Frontiers in Medicine, 12, 1-15.

[42] Budnarowski, D., Krawczyk, P., & Malinowski, T. (2025). AI and VR-driven adaptive educational environments. Applied Sciences, 15(16), 9067.

[43] Wen, J., Zhu, Y., Li, J., Zhu, M., & Tang, Z. (2025). Efficient multimodal transformer optimization for edge systems. IEEE Robotics and Automation Letters, 10(2), 1-9.

[44] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Large-scale multimodal visual instruction learning. NeurIPS 2023, 1-14.

[45] Ferrando, J., Sarti, G., Bisazza, A., & Costa-jussà, M. R. (2024). Explainability and optimization of transformer-based systems. ACM Computing Surveys, 57(3), 1-34.

[46] Ikram, Z. (2024, May). Dual-Domain Face Anti-Spoofing with Integrated Spatial and Frequency Analysis Neural Network. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 228-232). IEEE.

[47] Ma, Y., Song, Z., Zhuang, Y., Hao, J., & King, I. (2025). Vision-language-action systems for adaptive AI environments. ACM Computing Surveys, 58(1), 1-39.

[48] Farhah, N. S., Adnan, M., Alqarni, A. A., Uddin, M. I., & Aldhyani, T. H. H. (2026). Personalized multimodal adaptive learning using AI frameworks. IEEE Access, 14, 1-21.

[49] Xu, P., Zhu, X., & Clifton, D. A. (2023). Transformer architectures for multimodal learning systems. IEEE TPAMI, 45(9), 11213-11236.

[50] Ikram, Z. (2024, May). Hybrid deep neural network for face liveness detection in real-time video. In 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST) (pp. 188-193). IEEE.

[51] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., & Chen, X. (2023). Robotics transformer models for multimodal intelligent control. Conference on Robot Learning Proceedings, 1-15.

[52] Black, K., Brown, N., Driess, D., Esmail, A., & Equi, M. (2024). General multimodal flow models for adaptive decision systems. arXiv Preprint arXiv:2410.24164.

[53] Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., & Wang, X. (2024). Transferable vision-language multimodal architectures. arXiv Preprint arXiv:2410.05779.