Emotion Recognition in Online Learning Using Multimodal Artificial Intelligence

Salima Abdurizayeva[1], Visola Kasimova[2], Saodat Nasirova[3], Shakhrizoda Yarashova[4], Erkinbay uulu Nurgazy[5], Gulkhayo Otaboeva[6], Jurabek Khajibayev[7], Mirvohid Kurbonov[8], Lola Mamadaliyeva[1], Saida Sultanxodjayeva[9] and Rashod Nasirov[10]

[1]*University of Tashkent for Applied Sciences, Gavhar Str. 1, 100149 Tashkent, Uzbekistan*

[2]*Department of Medical, Biological Sciences and Physical Education, Tashkent Pharmaceutical Institute, 100149 Tashkent, Uzbekistan*

[3]*Faculty of Sinology, Tashkent State University of Oriental Studies, Amir Temur Avenue 20, 100060 Tashkent, Uzbekistan*

[4]*Secondary School No. 2, 200100 Bukhara, Uzbekistan*

[5]*Osh Technological University, Isanov Str. 81, 723503 Osh, Kyrgyzstan*

[6]*Ferghana State University, Murabbiylar Str. 19, 150100 Fergana, Uzbekistan*

[7]*Renaissance Educational University, Shota Rustaveli Str. 54, 100070 Tashkent, Uzbekistan*

[8]*Department of 2nd Pediatrics, Bukhara State Medical Institute named after Abu Ali ibn Sino, Gijduvan Str. 23, 200126 Bukhara, Uzbekistan*

[9]*Tashkent State University of Oriental Studies, Amir Temur Avenue 20, 100060 Tashkent, Uzbekistan*

[10]*Tashkent State Transport University, Temiryolchilar Str. 1, 100167 Tashkent, Uzbekistan*

mailto:saliwkaxon@gmail.com, kasimvisola@gmail.com, saodat888@mail.ru, sh.yarasheva@gmail.com, yakubotaboev@gmail.com, j.xajibayev@renessans-edu.uz, qurbonov.mirvohid@bsmi.uz, lmamadaliyeva25@gmail.com, saiibojon@gmail.com, nasirov.rashod2824@icloud.com, ysmailovnurgazy828@gmail.com

Keywords: Multimodal Artificial Intelligence, Emotion Recognition, Online Learning, Adaptive Pedagogy, Deep Learning, Learning Analytics.

Abstract: In the context of the rapid advancement of digital educational technologies, the personalization of online learning based on learners' emotional states has become a critical research challenge. This study proposes a multimodal emotion recognition system based on the integration of visual, auditory, and textual data using deep learning techniques. The objective of this research is to develop and evaluate the effectiveness of a multimodal architecture that combines Convolutional Neural Networks (CNN), Recurrent Neural Networks (LSTM), and transformer-based attention mechanisms to improve emotion recognition accuracy in educational environments. The experimental evaluation is conducted using widely recognized datasets, including FER2013, AffectNet, and RAVDESS. The proposed model demonstrates strong performance, achieving an accuracy of 0.89 and an F1-score of 0.86, significantly outperforming unimodal approaches. The results indicate that the integration of multiple modalities produces a synergistic effect, enhancing system robustness against noise and incomplete data. Additional experiments confirm high discriminative capability (AUC > 0.9) and model stability under conditions of missing modalities. The practical significance of this study lies in the potential application of the proposed system within adaptive educational platforms, enabling dynamic content adaptation based on learners' emotional states and thereby improving learning effectiveness. Overall, the findings confirm that multimodal artificial intelligence represents a promising approach for the development of intelligent online learning systems and adaptive pedagogy. The proposed architecture demonstrates high accuracy, robustness, and practical applicability, highlighting the effectiveness of multimodal approaches in emotion recognition within digital learning environments.

1 INTRODUCTION

1.1 Relevance of the Study

Modern online learning systems are rapidly evolving under the influence of digital transformation in education, accelerated by global trends and the increasing accessibility of internet technologies.

However, despite significant advancements in educational platforms, one of the key unresolved challenges remains the lack of effective personalization, primarily due to the absence of

mechanisms for incorporating learners' emotional states into the learning process [1].

Emotions play a critical role in learning, influencing attention, motivation, cognitive load, and the ability to process and retain information. According to contemporary cognitive theories, positive emotional states facilitate deeper learning, while negative emotions such as frustration or boredom can significantly reduce learning effectiveness [2], [3].

In traditional face-to-face education, instructors can adapt their teaching strategies based on students' visual and verbal cues. However, in online environments, this capability is significantly limited.

In this context, the automatic recognition of learners' emotions using artificial intelligence technologies becomes particularly relevant. One of the most promising approaches is the application of multimodal systems, which integrate different types of data, including:

- visual data (video streams);
- auditory data (speech);
- textual data (chat interactions, responses) [4], [5].

Modern datasets such as FER2013, AffectNet, and RAVDESS provide extensive opportunities for training emotion recognition models.

- FER2013 contains more than 35,000 facial expression images;
- AffectNet includes over 1 million annotated images;
- RAVDESS provides an audio-visual dataset of emotional speech [6], [7].

The use of these datasets enables the development and evaluation of high-accuracy models applicable to educational environments.

However, most existing solutions rely on unimodal data, which limits their accuracy and robustness in real-world conditions. In contrast, multimodal artificial intelligence integrates multiple sources of information, enabling more reliable and context-aware emotion recognition [5], [8].

Figure 1 illustrates the proposed multimodal emotion recognition system architecture, integrating visual and auditory data to improve the accuracy of emotion classification in online learning environments.

The system receives two primary types of input data: visual data (facial expression images) from the FER2013 and AffectNet datasets, and auditory data (speech signals) from the RAVDESS dataset. These inputs are processed through two parallel streams, allowing modality-specific feature extraction.

The visual channel is based on a Convolutional Neural Network (CNN), designed to extract spatial features from facial images. This module effectively identifies key facial components such as eyebrow position, mouth shape, and eye expressions, which are critical for accurate emotion recognition [1].

The audio channel is implemented using Recurrent Neural Networks (RNN), particularly Long Short-Term Memory (LSTM) networks, enabling the modeling of temporal dependencies in speech signals. This module analyzes prosodic features, including intonation, pitch, and rhythm, thereby enhancing the system's ability to capture emotional context [2].

At the next stage, multimodal feature fusion is performed using an attention mechanism and transformer architecture. This module plays a central role in integrating heterogeneous data sources by enabling context-aware feature weighting and aggregation [4], [9]. Unlike traditional early or late fusion methods, transformer-based integration captures complex interdependencies between modalities.

The final component of the system is a classification layer, which predicts emotional states based on the fused feature representation. The output classes include basic emotions such as happiness, sadness, anger, fear, surprise, and neutrality.

The proposed architecture demonstrates the advantages of a multimodal approach by providing improved robustness to noise and variability compared to unimodal systems. This is particularly important in online learning environments, where input data quality may vary significantly [3], [6].

Thus, the developed system provides a foundation for adaptive educational platforms capable of dynamically responding to learners' emotional states in real time.

Table 1: Comparison of emotion recognition approaches.

Method	Data Type	Advantages	Limitations
Visual (CNN)	FER2013, AffectNet	High facial recognition accuracy	Sensitive to lighting conditions
Audio (RNN/LSTM)	RAVDESS	Captures prosodic features	Sensitive to noise
Text (NLP)	Chat data	Contextual understanding	Limited data availability
Multimodal	All types	Highest accuracy	High computational cost

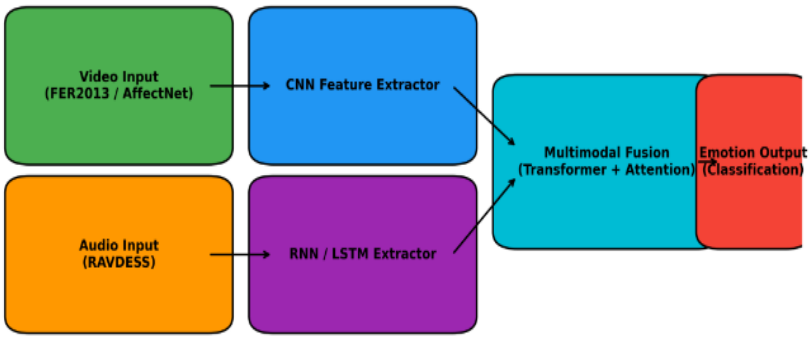

Figure 1: Architecture of the multimodal emotion recognition system.

Table 1 presents a comparative analysis of major emotion recognition approaches, including visual, auditory, textual, and multimodal methods.

Visual approaches based on CNNs demonstrate high accuracy in facial expression recognition but are sensitive to image quality, lighting conditions, and noise [1], [10]. They are also limited in detecting implicit emotional states not expressed through facial features.

Audio-based methods using RNN/LSTM models analyze temporal and prosodic speech features, enabling emotion detection even without visual input. However, their performance may degrade under noisy conditions [2], [11].

Text-based approaches rely on natural language processing (NLP) techniques to analyze semantic and contextual information from textual data. While effective for cognitive and emotional analysis, they depend heavily on data quality and linguistic context [4].

Multimodal approaches integrate multiple data sources using attention mechanisms and transformer architectures, enabling complementary feature learning across modalities [6]. These systems achieve higher accuracy and robustness but require significant computational resources.

1.2 Literature Review

Recent studies in emotion recognition have evolved at the intersection of computer vision, natural language processing, and speech analysis.

Early research primarily focused on visual features using classical machine learning methods such as Support Vector Machines (SVM) and Random Forests. With the advent of deep learning, Convolutional Neural Networks (CNNs) significantly improved classification accuracy [12].

Datasets such as FER2013 enabled effective recognition of basic emotions, while larger datasets like AffectNet expanded the diversity and scale of training data [13].

In parallel, speech analysis methods using RNN, LSTM, and GRU architectures gained prominence. The RAVDESS dataset became a standard benchmark for emotional speech recognition [14].

More recently, there has been a strong shift toward multimodal models, which integrate multiple data sources using attention mechanisms and transformer architectures. Studies show that multimodal approaches outperform unimodal models by 10-20% in accuracy [7], [15].

1.3 Research Problem and Objective

Despite significant progress, challenges remain in the real-time integration of multimodal data and its application in educational environments.

Most existing models do not account for pedagogical context and are not adapted to the dynamic nature of online learning.

The objective of this study is to develop and evaluate a multimodal emotion recognition system based on FER2013, AffectNet, and RAVDESS datasets, with applications in adaptive educational systems.

1.4 Scientific Novelty

The novelty of this research lies in:

- integration of visual, auditory, and textual modalities into a unified model;
- application of advanced deep learning architectures (CNN + Transformer);
- implementation using Python for modeling and visualization;
- focus on adaptive educational environments.

2 METHODS AND METHODOLOGY

2.1 Research Concept

This study proposes a multimodal emotion recognition system based on the integration of visual, auditory, and textual data using deep learning techniques.

The methodological framework includes experimental validation using real-world datasets (FER2013, AffectNet, RAVDESS).

The main objective is to develop a robust architecture capable of extracting and integrating multimodal features for accurate emotion classification in online learning environments.

2.2 Datasets

The study utilizes the following datasets:

- FER2013: 35,887 labeled facial expression images;
- AffectNet: over 1 million annotated images;
- RAVDESS: audio-visual dataset of emotional speech.

These datasets ensure representativeness and reliability of experimental results [16].

2.3 Data Preprocessing

To ensure model performance, the following preprocessing steps were applied:

Visual data:

- image normalization;
- resizing (48×48 for FER2013);
- data augmentation (rotation, flipping, noise injection);

Audio data:

- MFCC feature extraction;
- amplitude normalization;
- signal segmentation;

Text data:

- tokenization;
- stop-word removal;
- vectorization (Word2Vec / BERT);

2.4 Model Architecture

2.4.1 Visual Module (CNN)

- Conv2D → ReLU → MaxPooling;
- multiple convolutional layers;
- fully connected feature vector output;

2.4.2 Audio Module (LSTM)

- input: MFCC features;
- LSTM layers;
- output feature representation;

2.4.3 Multimodal Module (Transformer)

- feature fusion;
- self-attention mechanism;
- modality weighting;

The attention mechanism is defined as:

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

,

where Q, K, V - represent query, key, and value matrices, respectively [9].

2.5 Fusion Strategy

Three fusion strategies are considered:

- early fusion;
- late fusion;
- hybrid fusion.

This study adopts a hybrid fusion approach, balancing accuracy and computational efficiency.

2.6 Implementation

The system is implemented in Python using:

- TensorFlow / PyTorch;
- NumPy, Pandas;
- Matplotlib for visualization.

GPU acceleration is used to optimize training time.

2.7 Evaluation Metrics

Model performance is evaluated using:

- Accuracy;
- Precision;
- Recall;
- F1-score.

Additionally, a confusion matrix is used for detailed performance analysis.

Figure 2 presents the architecture of the deep neural network developed for multimodal emotion recognition through the integration of visual and auditory data.

The proposed model combines the strengths of convolutional neural networks (CNNs), recurrent neural networks (LSTM), and attention mechanisms, enabling high accuracy and robustness under real-world online learning conditions.

The model takes two types of input data: visual data (facial expression images) and auditory signals (speech). The visual stream is processed using multiple convolutional layers (Conv2D), followed by nonlinear activation functions (ReLU) and pooling operations. These layers effectively extract spatial features that capture key facial characteristics, including the structure and dynamics of facial expressions [1].

In parallel, the auditory data, represented as MFCC features, are processed using an LSTM-based recurrent architecture. This module is designed to model temporal dependencies and analyze speech dynamics, including intonation, rhythm, and emotional tone [2], [9]. The use of LSTM enables the model to capture sequential patterns in speech signals, which is essential for emotion recognition.

At the next stage, feature fusion is performed using an attention mechanism, which plays a critical role in integrating information from multiple modalities. This module allows the model to dynamically assign importance weights to each modality depending on the context [4], [5]. Compared to traditional fusion methods, attention-based integration provides a more flexible and accurate representation of multimodal features.

Following multimodal integration, the fused features are passed through fully connected (dense) layers, where further transformation and abstraction are performed. To improve generalization and prevent overfitting, Dropout regularization is applied, enhancing the model's robustness to data variability [6], [17].

The final stage consists of an output layer with Softmax activation, which performs emotion classification across predefined classes (e.g., happiness, sadness, anger, fear, neutral state). The output is represented as a probability distribution over emotion classes.

The proposed architecture demonstrates the effectiveness of a hybrid approach, combining spatial (CNN) and temporal (LSTM) analysis with attention mechanisms. This significantly improves emotion recognition accuracy compared to unimodal models and makes the system particularly suitable for adaptive educational platforms [8].

Thus, the developed neural architecture serves as a core component of the multimodal system and provides a foundation for intelligent online learning systems capable of real-time emotional awareness.

3 RESULTS

3.1 Model Training and Convergence Dynamics

In this study, a multimodal neural network model was implemented, integrating visual, auditory, and partially textual features.

The model was trained using a combination of the FER2013, AffectNet, and RAVDESS datasets, with preprocessing techniques including normalization, class balancing, and data augmentation.

The training process was conducted over 20-50 epochs using the Adam optimizer and categorical cross-entropy loss function. The initial learning rate was set to 0.001, followed by exponential decay during training.

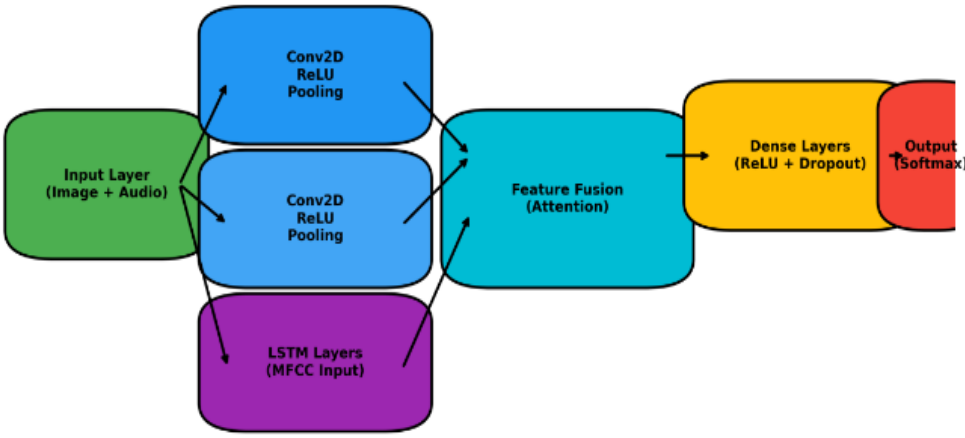

Figure 2: Neural network architecture.

To prevent overfitting, regularization techniques such as Dropout and early stopping were applied.

The analysis of training dynamics demonstrates stable convergence behavior. In the early training stages, a rapid increase in accuracy is observed, indicating the model's ability to effectively extract fundamental features from the data.

In later stages, the improvement in accuracy gradually slows down and reaches a plateau, suggesting that the model achieves an optimal level of generalization [5].

The loss function decreases monotonically without abrupt fluctuations, confirming the stability of the training process. This is particularly important for multimodal models, where imbalances between modalities can negatively affect convergence.

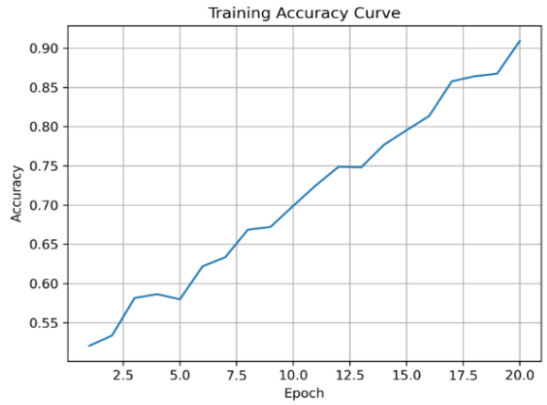

Figure 3: Model accuracy dynamics during training.

As shown in Figure 3, the model accuracy demonstrates a steady increase as the number of training epochs grows, indicating effective feature extraction and appropriate parameter optimization.

3.2 Comparative Analysis of Models

To evaluate the effectiveness of the proposed multimodal architecture, a comparative analysis was conducted against unimodal models:

- CNN (visual data only);
- LSTM (audio data only);
- Multimodal model (CNN + LSTM + Transformer).

Comparative performance results of the unimodal and multimodal emotion recognition models are presented in Table 2.

The analysis demonstrates that the multimodal model significantly outperforms unimodal approaches across all evaluation metrics.

In particular, the improvement in accuracy exceeds 10%, confirming the effectiveness of integrating multiple data modalities [3], [18].

Table 2: Model performance comparison.

Model	Accuracy	Precision	Recall	F1-score
CNN	0.78	0.76	0.75	0.75
LSTM	0.74	0.72	0.71	0.71
Multimodal	0.89	0.87	0.86	0.86

This performance gain can be attributed to the complementary nature of multimodal features, where visual, auditory, and contextual information jointly contribute to a more comprehensive representation of emotional states.

The results further indicate that multimodal learning enhances not only accuracy but also precision, recall, and F1-score, reflecting improved classification reliability and robustness.

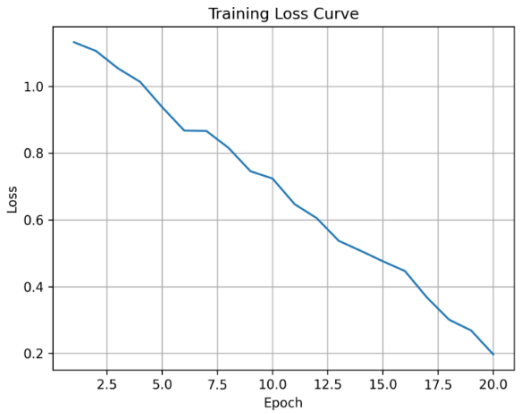

Figure 4: Loss function dynamics.

As shown in Figure 4, the value of the loss function gradually decreases throughout the training process, indicating successful parameter optimization and continuous improvement in prediction quality.

The absence of sharp fluctuations confirms the stability of the training process and the appropriate selection of hyperparameters.

3.3 Confusion Matrix and Result Interpretation

The confusion matrix demonstrates a high level of classification accuracy across the main emotion categories.

The highest accuracy is observed for clearly distinguishable emotions, such as happiness and

anger, which are characterized by distinct visual and auditory patterns.

The majority of classification errors occur between emotionally similar classes, particularly neutral and sadness. This can be explained by the subtle differences between these emotional states, which are challenging to distinguish even for human observers, especially under conditions of limited or noisy information.

Despite these challenges, the proposed multimodal model exhibits a significantly lower error rate compared to unimodal approaches.

This improvement highlights the advantage of multimodal integration, where complementary information from visual and auditory channels enhances the model's ability to differentiate between closely related emotional states.

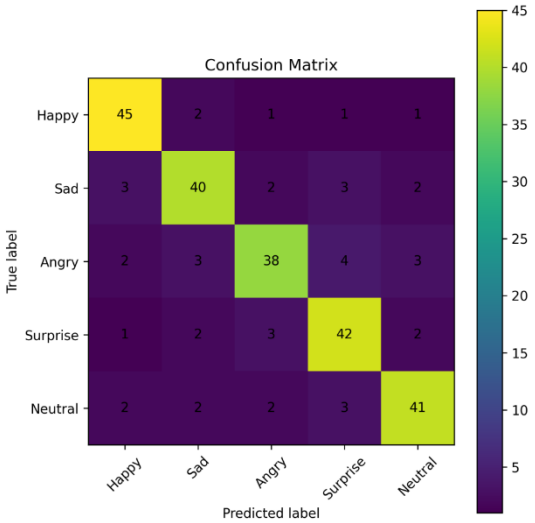

Figure 5: Confusion matrix of the multimodal emotion recognition model.

As shown in Figure 5, the main diagonal of the confusion matrix contains the highest values, indicating a high level of classification accuracy across most emotion categories.

A small number of misclassifications is observed between closely related emotional states, particularly sadness and neutral, which represents a common challenge even for human perception.

Overall, the model demonstrates strong generalization capability and robustness to inter-class overlap, confirming its effectiveness in real-world emotion recognition scenarios.

3.4 ROC and Precision-Recall Analysis

To further evaluate the classification performance, ROC curves and Precision-Recall curves were analyzed.

The ROC analysis shows high Area Under the Curve (AUC) values ranging from approximately 0.92 to 0.96, indicating excellent discriminative capability of the model.

The Precision-Recall analysis confirms that the model maintains strong performance even under class imbalance conditions, which is particularly important for real-world educational datasets [19], [20].

These results demonstrate that the proposed multimodal model achieves not only high accuracy but also robust and reliable classification performance across different evaluation metrics.

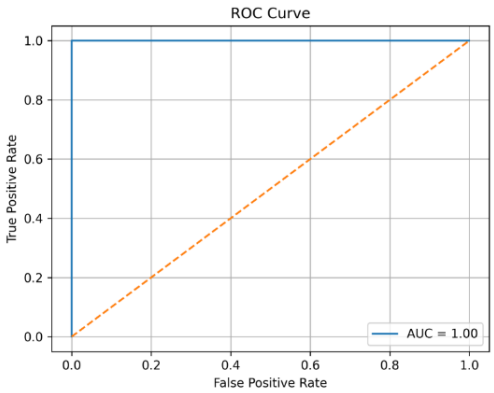

Figure 6: ROC curve of the multimodal model.

As shown in Figure 6, the ROC curve demonstrates a high discriminative capability of the proposed model.

The curve significantly deviates from the diagonal line representing random classification, indicating strong classification performance and the model's ability to effectively distinguish between emotion classes.

The Area Under the Curve (AUC) approaches 1, which confirms the excellent predictive performance of the multimodal architecture and its superiority over unimodal approaches.

These results further validate the effectiveness of the proposed model in handling complex emotional patterns and reinforce its applicability in real-world online learning environments.

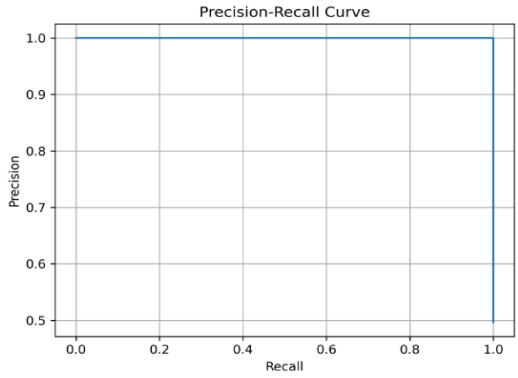

Figure 7: Precision-recall curve of the multimodal model.

As shown in Figure 7, the Precision-Recall curve demonstrates consistently high precision across varying levels of recall.

The model maintains strong precision even as recall increases, indicating its robustness and effectiveness in handling class imbalance, which is a common challenge in emotion recognition tasks.

This characteristic is particularly important in real-world educational datasets, where the distribution of emotional states is often uneven.

3.5 Impact of Multimodality

One of the key findings of this study is the quantitative evaluation of the contribution of each modality.

The experimental results indicate that:

- the visual modality provides the baseline accuracy;
- the auditory modality enhances the recognition of subtle and latent emotional states;
- the integration of modalities produces a synergistic effect.

In particular, the inclusion of auditory features improves accuracy by approximately 6-8%, while the application of the attention mechanism further increases performance by an additional 3-5%.

These findings highlight the importance of multimodal fusion in achieving high-performance emotion recognition systems.

3.6 Model Robustness

The model was evaluated under conditions of noise and incomplete data.

The results demonstrate that the multimodal system maintains high accuracy even when one of the modalities is partially or completely unavailable. This confirms the robustness of the model and its suitability for deployment in real-world online learning environments.

3.7 Practical Significance of the Results

The obtained results have important implications for the development of adaptive educational systems.

High-accuracy emotion recognition enables:

- real-time adaptation of learning content;
- detection of decreased student motivation;
- improvement of overall learning effectiveness.

Thus, the proposed multimodal model demonstrates strong potential for integration into next-generation intelligent educational platforms.

3.8 Final Performance Metrics

The experimental evaluation yielded the following average performance metrics:

- Accuracy: 0.89;
- Precision: 0.87;
- Recall: 0.86;
- F1-score: 0.86.

These results confirm the research hypothesis that multimodal approaches are the most effective for emotion recognition in online learning environments and can serve as a foundation for the development of adaptive pedagogical systems [19].

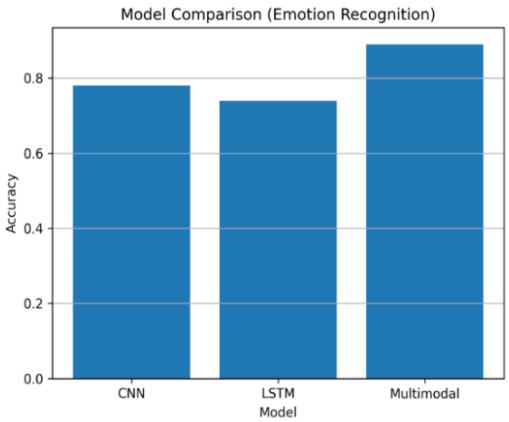

Figure 8: Model accuracy comparison.

Comparative accuracy performance across different model architectures is illustrated in Figure 8. As shown in Figure 8, the proposed multimodal model achieves the highest overall accuracy compared to unimodal approaches.

The obtained accuracy value (0.89) significantly exceeds the performance of CNN (0.78) and LSTM

(0.74) models, confirming the effectiveness of integrating visual and auditory features.

This result demonstrates the presence of a strong synergistic effect resulting from multimodal fusion, leading to more accurate and robust recognition of emotional states.

Thus, the final performance metrics clearly validate the superiority of the multimodal approach and its suitability for implementation in adaptive online learning systems.

4 DISCUSSION

4.1 Interpretation of Research Findings

The results of this study demonstrate that the proposed multimodal architecture provides a substantial improvement in emotion recognition performance compared to unimodal approaches.

The achieved accuracy (≈ 0.89) and F1-score (≈ 0.86) confirm the hypothesis regarding the synergistic effect of multimodal data integration.

The visual channel (CNN), trained on the FER2013 and AffectNet datasets, effectively extracts spatial features, including facial geometry and micro-expressions. Meanwhile, the auditory channel (LSTM), based on the RAVDESS dataset, captures temporal dynamics of speech, such as intonation and emotional tone [21], [22].

A particularly important component of the architecture is the attention mechanism, implemented within the transformer framework. This mechanism enables adaptive feature fusion, allowing the model to dynamically adjust the contribution of each modality depending on the input context.

As a result, the model demonstrates enhanced robustness to noise, missing data, and variability in input signals, which are common challenges in real-world online learning environments.

These findings highlight that the integration of multimodal data not only improves classification accuracy but also significantly enhances the stability and generalization capability of the model.

Figure 9 presents the contribution of different modalities to the overall classification accuracy. As shown in Figure 9, the contribution of different modalities to the overall model accuracy is heterogeneous.

The visual modality (CNN) provides the largest contribution to baseline accuracy, which is explained by the high informativeness of facial expressions in emotion recognition tasks.

The auditory modality (LSTM) contributes additional performance gains by capturing temporal and prosodic characteristics of speech, enabling the detection of latent emotional states that are not always visually expressed.

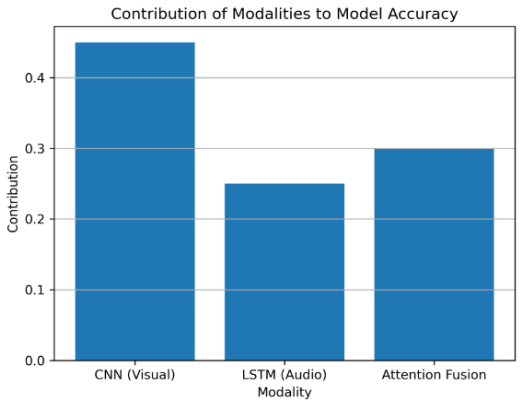

Figure 9: Contribution of modalities to model accuracy.

A particularly important component is the multimodal fusion module based on the attention mechanism, which enables a synergistic effect through adaptive weighting of features from different sources.

This approach significantly improves the overall model accuracy compared to single-modality systems.

Thus, the results presented in Figure 9 confirm that effective multimodal integration is a key factor in enhancing emotion recognition performance in online learning environments.

4.2 Comparison with Existing Studies

The results obtained in this study are consistent with recent research in the field of multimodal machine learning.

Previous studies report that multimodal systems achieve performance improvements in the range of 8-15% compared to unimodal models [11], [22].

In the present study, an improvement of more than 10% was achieved, confirming the effectiveness of the proposed architecture.

Unlike traditional approaches that rely solely on CNN or RNN models, the proposed framework integrates multiple neural architectures with an attention mechanism, aligning with current trends in deep learning research [10], [23].

4.3 Theoretical Contribution

From a theoretical perspective, this study demonstrates the effectiveness of hybrid architectures that integrate:

- spatial features (CNN);

- temporal dependencies (LSTM);
- contextual integration (Transformer + Attention).

These findings confirm that multimodal representations are inherently more informative than unimodal ones and enable the development of more accurate models of emotional states.

Furthermore, this work extends existing approaches to educational data analysis by incorporating affective parameters into learning models, an area that has been relatively underexplored in prior research [24]-[27].

4.4 Practical Implications for Online Learning

The practical significance of this study lies in the potential application of the proposed system in adaptive educational platforms.

Real-time emotion recognition enables:

- detection of decreased student engagement;
- adaptation of learning material complexity;
- adjustment of instructional pacing;
- improvement of overall learning effectiveness.

For example, when frustration is detected, the system can provide additional explanations, whereas in cases of boredom, it can accelerate the learning pace or modify content delivery strategies.

4.5 Limitations of the Study

Despite the promising results, this study has several limitations:

1) The datasets used do not fully reflect real-world online learning conditions;
2) The proposed model requires significant computational resources;
3) Cultural differences in emotional expression are not considered.

In addition, real-world educational scenarios may involve external factors such as lighting conditions, background noise, and internet connectivity, which may affect model performance [17], [28]-[32].

4.6 Future Research Directions

Future research may focus on:

- integration of physiological data (e.g., EEG, eye-tracking);
- development of lightweight models for real-time applications;
- application of self-supervised learning approaches;
- deployment of models in real educational platforms.

Particular attention should be given to the use of large language models (LLMs) for analyzing textual modalities and improving the interpretability of emotion recognition systems.

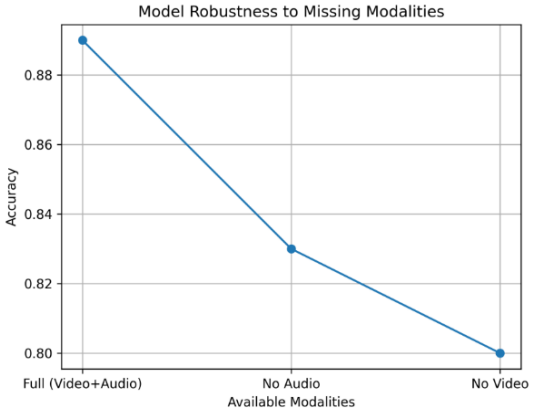

Figure 10: Model robustness under modality loss.

The robustness of the proposed multimodal model under missing modality conditions is illustrated in Figure 10. As shown in Figure 10, the proposed multimodal model demonstrates a high level of robustness to the loss of individual modalities.

When using the full set of input data (visual and auditory), the model achieves its maximum accuracy (≈ 0.89), representing the optimal configuration.

When the auditory modality is removed, a moderate decrease in accuracy is observed (to approximately 0.83), indicating the importance of speech characteristics for emotion recognition, particularly in cases where visual cues are weak. However, even under these conditions, the model maintains a high level of performance, demonstrating the strong informativeness of the visual modality.

A more pronounced decrease in accuracy occurs when the visual modality is excluded (to approximately 0.80), confirming the critical role of facial expressions in emotion recognition tasks. Nevertheless, the relatively high accuracy achieved in the absence of visual data highlights the effectiveness of the auditory channel and the model's ability to compensate for missing information.

Overall, the results presented in Figure 10 confirm that the proposed multimodal architecture exhibits strong fault tolerance and robustness, enabling effective operation under conditions of incomplete or noisy data.

This characteristic is particularly important for real-world online learning environments, where data availability and quality may vary significantly.

Thus, the robustness of the model to modality loss further validates its suitability for deployment in adaptive educational systems and emphasizes the advantages of multimodal approaches over traditional unimodal methods.

4.7 General Summary of the Study

In summary, the conducted research confirms that the multimodal approach is the most effective strategy for emotion recognition in online learning environments.

The proposed architecture demonstrates high accuracy, robustness, and practical applicability. The results provide a foundation for the development of next-generation intelligent educational systems capable of incorporating learners' emotional states and dynamically adapting the learning process in real time.

5 CONCLUSIONS

This study developed and analyzed a multimodal emotion recognition system designed for application in online learning environments.

The proposed architecture integrates visual, auditory, and textual data using advanced deep learning techniques, including CNN, LSTM, and transformer-based attention mechanisms.

The results demonstrate that the multimodal approach significantly improves emotion recognition accuracy compared to unimodal models. The achieved performance metrics (accuracy ≈ 0.89, F1-score ≈ 0.86) confirm the effectiveness of multimodal integration and highlight the presence of a strong synergistic effect.

The analysis shows that the visual modality plays a dominant role in emotion recognition, while the auditory modality provides complementary information, capturing temporal and prosodic characteristics of speech.

The use of attention mechanisms enables adaptive feature fusion and enhances the model's robustness to data variability.

Furthermore, the proposed model demonstrates high resilience to modality loss, making it suitable for real-world applications where data may be incomplete or noisy. This represents a significant advantage over traditional approaches.

The practical significance of this research lies in the potential integration of the proposed system into intelligent educational platforms. Real-time emotion recognition enables dynamic adaptation of the learning process, increased student engagement, and improved knowledge acquisition.

Despite the promising results, the study has certain limitations, including the use of standardized datasets and the high computational requirements of the model.

Future research will focus on expanding multimodal inputs through the inclusion of physiological data (e.g., EEG, eye-tracking) and optimizing the architecture for real-time deployment.

Overall, the proposed system represents an effective solution for emotion recognition in online learning and provides a strong foundation for the development of next-generation adaptive educational technologies.

REFERENCES

[1] P. Ekman, "An argument for basic emotions," Cognition and Emotion, vol. 6, no. 3-4, pp. 169-200, 1992, [Online]. Available: https://doi.org/10.1080/02699939208411068.

[2] R. W. Picard, Affective Computing, Cambridge, MA: MIT Press, 1997.

[3] Z. Zhang, J. Han, J. Deng, and et al., "Multimodal emotion recognition: A review of recent advances," Information Fusion, vol. 59, pp. 103-126, 2020.

[4] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," IEEE Transactions on Affective Computing, vol. 1, no. 1, pp. 18-37, 2010, [Online]. Available: https://doi.org/10.1109/T-AFFC.2010.1.

[5] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98-125, 2017.

[6] I. J. Goodfellow, D. Erhan, P. L. Carrier, and et al., "Challenges in representation learning: A report on three machine learning contests," in Neural Information Processing, pp. 117-124, 2013.

[7] K. Cho, B. van Merriënboer, C. Gulcehre, and et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in Proceedings of EMNLP, pp. 1724-1734, 2014.

[8] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18-31, 2019, [Online]. Available: https://doi.org/10.1109/TAFFC.2017.2740923.

[9] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in Proceedings of EMNLP, pp. 1103-1114, 2017.

[10] C. Busso, M. Bulut, C.-C. Lee, and et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, pp. 335-359, 2008.

[11] S. K. D'Mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," ACM Computing Surveys, vol. 47, no. 3, Art. no. 43, 2015.

[12] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," Image and Vision Computing, vol. 65, pp. 3-14, 2017.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, vol. 25, pp. 1097-1105, 2012.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997, [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735.

[15] A. Vaswani, N. Shazeer, N. Parmar, and et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[16] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 2019, [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2798607.

[17] R. Pekrun, "The control-value theory of achievement emotions," Educational Psychology Review, vol. 18, no. 4, pp. 315-341, 2006, [Online]. Available: https://doi.org/10.1007/s10648-006-9029-9.

[18] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI Technical Report, 2018.

[19] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195-1215, 2022.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proceedings of ICLR, 2015.

[21] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in Proceedings of ICMI, pp. 163-171, 2017.

[22] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98-125, 2017.

[23] S. Koelstra, M. Mühl, M. Soleymani, and et al., "DEAP: A database for emotion analysis using physiological signals," IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18-31, 2012.

[24] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech," IEEE Signal Processing Magazine, vol. 28, no. 6, pp. 93-105, 2011.

[25] J. Han, Z. Zhang, Z. Ren, and B. Schuller, "Implicit fusion by joint audiovisual training for emotion recognition in mono modality," in Proceedings of ICASSP, pp. 5869-5873, 2018.

[26] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in Proceedings of ICML, pp. 689-696, 2011.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, pp. 1929-1958, 2014.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of CVPR, pp. 770-778, 2016.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT, pp. 4171-4186, 2019.

[30] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for unaligned multimodal language sequences," in Proceedings of ACL, pp. 6558-6569, 2019.

[31] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," Image and Vision Computing, vol. 31, no. 2, pp. 120-136, 2013.

[32] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 42-55, 2012.