Tom 16/2024, ss. 359-372
ISSN 0860-5637
e-ISSN 2657-7704
DOI: 10.19251/rtnp/2024.16(17)
www.rtnp.mazowiecka.edu.pl

Oleksii Sysoiev
Akademia Mazowiecka w Płocku
ORCID: 0000-0001-5899-0244

Podejście multimodalne do analizy stanów emocjonalnych uczniów w systemach edukacyjnych opartych na sztucznej inteligencji: aspekty teoretyczne i metodologiczne

Multimodal Approach to Analysing Students' Emotional States in AI-Enabled Educational Systems: Theoretical and Methodological Aspects

Streszczenie
Artykuł przedstawia teoretyczne i metodologiczne podstawy podejścia multimodalnego do rozpoznawania emocji w systemach edukacyjnych wspieranych przez sztuczną inteligencję. Przeanalizowano współczesne rozwiązania, integrację modeli emocji oraz nauczanie online. Usystematyzowano podejścia na podstawie analizy wskaźników tekstowych, audiowizualnych i fizjologicznych. Celem badania jest określenie podstaw koncepcyjnych dla systemów adaptacyjnych uwzględniających stany emocjonalne uczniów.

Słowa kluczowe: analiza multimodalna, sztuczna inteligencja, systemy edukacyjne, stany emocjonalne, nauczanie adaptacyjne, rozpoznawanie emocji

Abstract
This article explores theoretical and methodological foundations of the multimodal approach to emotion recognition in AI-supported education. The study reviews current solutions, integration of emotion models, and online learning tools. Approaches are systematized by analysing textual, audiovisual, and physiological indicators. The aim is to define a conceptual basis for adaptive systems that can respond to students' emotional states.

Keywords: multimodal analysis, artificial intelligence, educational systems, emotional states, adaptive learning, emotion recognition

1. Introduction

Research relevance. Emotional analysis plays a crucial role in AI-enabled educational systems (AIEd). It helps identify students' emotional states, such as interest, confusion, or disappointment, allowing platforms to adapt the learning process to individual needs. AutoTutor studies conducted over 17 years [Nye, Graesser and Hu, 2014] have shown that considering emotional states can improve student performance by 20-30%. [D'Mello and Graesser, 2013] note that integrating emotional analysis into AIEd enhances cognitive outcomes and increases student engagement through adaptive emotional responsiveness. Modern developments in affective computing, initiated by Rosalind Picard [Picard, 1997], make such approaches not only possible but also in demand.

Multimodal approach in emotional analysis. Emotional analysis is the process of recognizing, interpreting, and classifying emotional states. It employs a multimodal approach that combines data from several channels: facial expression analysis (emotion recognition through facial features), voice analysis (assessment of intonation, volume, and timbre), text analysis (interpretation of lexical and syntactic features), and physiological indicators (collection of heart rate and electrodermal activity data).

The multimodal approach enables more accurate determination of students' emotional states by combining these data

sources. For example, the system simultaneously analyses facial expressions via webcam, voice intonations, emotional colouring of text responses, and biometric signals, creating a comprehensive understanding of the student's emotional state and allowing real-time adaptation of the learning process.

Challenges. Despite obvious advantages, the multimodal approach faces several challenges: technical issues (integration of heterogeneous data and high computational costs [Sharma and Papamitsiou, 2019; Gandhi et al., 2023], methodological difficulties (analysis of complex emotions and consideration of cultural differences [Das and Das, 2024], and ethical aspects (student data privacy [Cukurova and Giannakos, 2020]. These challenges necessitate the development of new approaches aimed at overcoming technical, methodological, and ethical barriers, which forms the basis for this study's goals and objectives.

Research goals and objectives. This study aims to create a theoretical and methodological foundation for multimodal emotion analysis in the development of adaptive educational systems.

Prospects. Implementation of the proposed approach will enable the creation of next-generation educational platforms that consider a wide spectrum of students' emotional states, increasing engagement and improving learning outcomes. This research represents an important step toward overcoming existing limitations of the multimodal approach to emotion analysis and ensuring its practical application in educational environments.

2. Review of Existing AIEd Solutions with Emotional Components

Modern artificial intelligence systems in education (AIEd) actively implement multimodal approaches to analyse students' emotions, enabling adaptation of the educational process to their individual needs. These approaches combine textual data, audio-visual signals, and physiological indicators such as heart rate and

electrodermal activity. Data processing is performed using deep learning, including architectures such as CNN and RNN, which provide high accuracy in emotion analysis. Cognitive models that account for the dynamics of emotional states also play an important role in developing adaptive systems. The results of emotional analysis are used to adapt educational content in real-time, address difficulties, and increase student engagement.

For example, research by Poria et al. [Poria et al.,2015] demonstrated that using multimodal data, such as text, audio, and video, combined with deep neural networks, achieves emotion classification accuracy of 87.95%. However, their approach requires significant computational resources and high-quality data, limiting its application in real-world conditions. Meanwhile, Cukurova et al. [Cukurova et al., 2019] demonstrated that integrating audio and video data achieves 78% accuracy in classifying social and emotional aspects of learning, which is 13% higher than unimodal methods. However, this approach faces computational load issues that limit its real-time implementation and requires high-quality data and substantial computational resources. As Cukurova [Cukurova, 2024] later notes, modern AIEd solutions primarily rely on inductive methods, allowing them to adapt to various educational scenarios. Nevertheless, such methodology may limit the development of new models due to their dependence on existing data.

Multimodal analysis has also proven effective in engineering disciplines. Maoz [Maoz, 2024] demonstrated that using multimodal methods, including visual data and speech analysis, can improve student performance by 15% compared to traditional methods. Meanwhile, Kim et al. [Kim et al., 2018] proposed integrating physiological sensors to monitor students' emotional reactions. Their approach increased student engagement by 20%, although its widespread implementation is limited by equipment costs.

The most promising results are achieved using combined models, such as in the work “Intelligence Unleashed: An Argument for AI in Education. Pearson Education” of Luckin Rose and Holmes Wayne [Luckin and Holmes,2016]. Their approach, based on integrating cognitive parameters and multimodal analysis, increased learning process satisfaction by 25%. However, such systems require further optimization to ensure scalability with large volumes of data.

Table 1: Key Approaches and Effectiveness Metrics of Multimodal Analysis in AIEd

Approach	Data Types	Processing Methods	Effectiveness Metrics	Limitations
Poria et al. (2015)	Text, audio	Deep learning	87.95% accuracy	Data quality dependency
Cukurova et al. (2019)	Audio, video	Multimodal analysis	78% accuracy	Computational costs
Maoz (2024)	Video, audio	Multimodal data integration	15% performance improvement	Large data processing
Kim et al. (2018)	Physiological	Sensor analysis	20% engagement increase	High equipment cost
Luckin & Holmes (2016)	Cognitive, physiological	Combined models	25% satisfaction increase	Scalability issues

Table prepared by the author

These studies confirm (Table 1) that multimodal approaches open new possibilities for creating adaptive educational systems. Nevertheless, important challenges remain, such as the need to optimize computational resources, ensure data privacy, and improve input data quality. Future AIEd development is linked to expanding the use of transformer architectures, which enable more efficient processing of multimodal data, and developing hybrid models capable of reducing costs and improving scalability.

3. Integration of Classical Emotion Models into the Digital Environment

The integration of classical emotion models into the digital educational environment represents a crucial step in developing adaptive AIEd systems. Ekman's basic emotions theory and Barrett's constructed emotions theory present two complementary approaches that provide a foundation for analysing students' emotional states. Their application in digital systems requires both methodological adaptation and corresponding technological solutions.

Ekman's theory of basic emotions [Ekman, 1992] focuses on universal emotions such as joy, fear, and anger. In digital systems, this theory is implemented through computer vision algorithms such as AffectNet, which analyse facial expressions and features. The updated AffectNet+ database achieves 90.4% accuracy in emotion classification using deep neural networks [Fard et al., 2024]. Research by Waldner Daniel and Mitra Siddhartha [Waldner and Mitra, 2024] using the ArcFace model demonstrates similar results with 87.6% accuracy in analysing eight emotion classes. In systems with limited computational resources, simpler models such as Scaleddown EfficientNet can be used, achieving 61.3% accuracy [Fontes, 2023]. This makes it suitable for smaller educational platforms where resources are constrained.

Barrett's approach [Barrett, 2017], based on the theory of constructed emotions, emphasizes the contextual and cultural variability of emotional states. For text data analysis, the BERT transformer architecture is widely used, enabling identification of emotional signals in texts with up to 85% accuracy [Poria et al., 2015]. These methods find application in educational platforms such as AutoTutor, where student engagement increases by 20% through content adaptation based on emotional data. However, analysis of textual and audio emotions requires significant computational resources and may face scalability issues.

Deductive approaches characteristic of Ekman's theory [Ekman, 1992] focus on testing predetermined hypotheses, such as predictable emotional reactions to successful task completion. Meanwhile, inductive methods inherent in Barrett's theory allow for identifying new emotion patterns by analysing large volumes of data. Cukurova's [Cukurova, 2024] research emphasizes the necessity of combining these approaches, enabling the creation of universal and adaptive educational technologies. Such hybrid systems, combining textual, audiovisual, and physiological analysis, can account for both stable and dynamic aspects of student emotions.

Table 2: Results of Emotion Analysis Model Applications

Model or Method	Accuracy	Application Area	Source
AffectNet+	90.4%	Facial expression analysis	Fard et al., 2024
ArcFace on AffectNet	87.6%	Emotion recognition	Waldner & Mitra, 2024
Scaleddown EfficientNet	61.3%	Resource-constrained systems	Fontes, 2023
EfficientNet (average accuracy)	85%	Facial expression recognition	Utami et al., 2022
Text emotion analysis (BERT)	85%	Text data in learning	Poria et al., 2015

Table prepared by the author

Table 2 illustrates the comparative analysis of contemporary emotion recognition models and their practical applications in educational environments. The data demonstrates the significant advancement in facial expression analysis technology, with AffectNet+ achieving remarkable accuracy of 90.4%, while text-based emotion analysis through BERT architecture maintains consistent performance at 85%. These findings underscore the varying effectiveness of different methodological approaches across application domains, reflecting the current state of emotion recognition technology in educational settings [Fard et al., 2024; Poria et al., 2015].

4. Features of Emotion Recognition in Online Learning

Emotion recognition in online learning serves as an essential tool for adaptive educational technologies, enabling consideration of students' emotional states to enhance their engagement and academic performance. Modern systems employ various methods, including facial expression analysis, voice analysis, text analysis, and their combinations in multimodal models. These approaches vary in accuracy and applicability depending on available data and platform technical capabilities.

Facial expression analysis methods based on computer vision algorithms have demonstrated high accuracy. For instance, research by Alsadoon Abeer and Singh Amit Kumar [Alsadoon and Singh, 2018] shows that emotion recognition accuracy based on facial expressions reaches 85%. However, such systems are sensitive to image quality and lighting conditions. Voice data analysis, as noted in the work of Bahreini et al. [Bahreini et al, 2016], improves accuracy when integrated with other modalities but also faces challenges with noise and requires high-quality audio input. Text analysis methods, such as BERT transformers, achieve 85% accuracy when processing students' written responses [Poria et al., 2015], but they are limited in cases of brief or ambiguous texts.

Multimodal systems combining facial expression, voice, and text data show the highest accuracy. For example, the system described by Gupta et al. [Gupta et al., 2023] achieves accuracy up to 90% through deep neural networks such as CNN and LSTM. However, such solutions require significant computational resources and complex algorithms for data synchronization. Nevertheless, they provide the most adaptive and personalized approach in online learning.

Despite significant progress in developing emotion recognition technologies, their implementation faces technical and ethical challenges. Collection and processing of emotional data require strict privacy protection protocols to comply with modern privacy standards. In the research of Yu et al. [Yu et al., 2024] emphasize

the importance of adhering to ethical norms when developing such systems, as well as the need for optimization for large-scale use. Successful implementation of these technologies requires an interdisciplinary approach combining achievements in computer science and pedagogy.

Table 3: Effectiveness and Limitations of Emotion Recognition Methods

Analysis Method	Accuracy	Key Limitations	Source
Facial Analysis	Up to 85%	Dependence on image quality	Alsadoon & Singh, 2018
Voice Data Analysis	Up to 80%	Vulnerability to noise, quality requirements	Bahreini et al., 2016
Text Analysis	Up to 85%	Limited with brief responses	Poria et al., 2015
Multimodal Systems	Up to 90%	High computational costs	Gupta et al., 2023

Table prepared by the author

Table 3 presents a comprehensive overview of emotion recognition methods' effectiveness and their inherent limitations in online learning environments. The data reveals that multimodal systems achieve the highest accuracy (90%), significantly outperforming single-modality approaches. This comparative analysis highlights the trade-offs between accuracy and implementation challenges, particularly regarding computational requirements and data quality dependencies [Gupta et al., 2023; Alsadoon & Singh, 2018].

5. Conclusion

The conducted analysis of theoretical and methodological foundations of the multimodal approach to emotional state analysis in AIEd systems allows us to draw several significant conclusions.

First, modern AIEd solutions demonstrate significant progress in multimodal emotion analysis. The integration of various data types–textual, audiovisual, and physiological–enables high

accuracy in recognizing students' emotional states. Systems using combined models, such as the solution by Gupta et al. [Gupta et al., 2023], demonstrate accuracy up to 90%, confirming the effectiveness of the multimodal approach.

Second, the integration of classical emotion theories into the digital environment creates a solid theoretical foundation for AIEd systems development. Combining Ekman's and Barrett's approaches allows consideration of both universal and contextual aspects of emotional states. Modern technical solutions, such as AffectNet+ with 90.4% accuracy [Fard et al., 2024], demonstrate the possibility of effectively implementing these theoretical models.

Third, the features of emotion recognition in online learning require a comprehensive approach to system development. The conducted analysis shows that successful implementation of multimodal solutions depends on balancing recognition accuracy, computational efficiency, and adherence to ethical norms.

However, the research has identified several significant challenges requiring further investigation:

1. Optimization of computational resources to ensure real-time operation
2. Development of privacy protection methods for emotional data collection
3. Creation of scalable solutions for widespread implementation in educational practice
4. Further development of this direction is envisioned in the following aspects:
5. Development of hybrid models combining efficiency and analysis accuracy
6. Creation of adaptive algorithms considering individual student characteristics
7. Integration of ethical principles into AIEd system development

Thus, the conducted research not only systematizes existing approaches to multimodal emotion analysis in education but

also forms a basis for further development of adaptive educational technologies. The presented results can serve as a theoretical foundation for developing the next generation of AIEd systems capable of effectively considering emotional aspects of learning.

Bibliography

Alsadoon Abeer, Singh Amit Kumar. 2018. "An emotion recognition model based on facial recognition in virtual learning environment." Procedia Computer Science. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050917327679 (accessed December 7, 2024).

Bahreini Kiavash, Nadolski Rob, Westera Wim. 2016. "Towards multimodal emotion recognition in e-learning environments." Interactive Learning Environments. Retrieved from https://research.ou.nl/files/1007887/Bahreini,%20Nadolski,%20and%20Westera%202014-1.pdf (accessed December 7, 2024).

Barrett Lisa Feldman. 2017. How Emotions Are Made: The Secret Life of the Brain. Houghton Mifflin Harcourt. Retrieved from https://books.google.com/books?id=kTQfDQAAQBAJ (accessed December 7, 2024).

Cukurova Mutlu. 2024. "A Syllogism for Designing Collaborative Learning Technologies in the Age of AI and Multimodal Data." Retrieved from https://discovery.ucl.ac.uk/id/eprint/10086137/ (accessed December 7, 2024).

Cukurova Mutlu, Giannakos Michail. 2020. "The Promise and Challenges of Multimodal Learning Analytics." Retrieved from https://discovery.ucl.ac.uk/id/eprint/10106543/1/Accepted_MMLA%20paper.pdf (accessed December 7, 2024).

Cukurova Mutlu, Kent Carmel, Luckin Rose. 2019. "The Value of Multimodal Data in Classification of Social and Emotional Aspects of Tutoring." Proceedings of AIED 2019, 454-458. Retrieved from https://discovery.ucl.ac.uk/id/eprint/10086137/1/AIED_2019_paper_175.pdf (accessed December 7, 2024).

Das Bijay, Das Rama. 2024. "A Systematic Review of Trimodal Affective Computing Approaches: Text, Audio, and Visual Integration in Emotion Recognition and Sentiment Analysis." Retrieved from https://www.sciencedirect.com/science/article/pii/S0957417424017196 (accessed December 7, 2024).

D'Mello Sidney K., Graesser Arthur C. 2013. "AutoTutor and Affective AutoTutor: Learning by Talking with Cognitively and Emotionally Intelligent Computers That Talk Back." ACM Transactions on Interactive Intelligent Systems. Retrieved from https://www.ida.liu.se/divisions/hcs/seminars/cogsciseminars/Papers/dmello-TiiS12.pdf (accessed December 7, 2024).

Ekman Paul. 1992. "An Argument for Basic Emotions." Cognition and Emotion 6(3-4): 169-200. Retrieved from https://www.researchgate.net/publication/222933508_An_argument_for_basic_emotions (accessed December 7, 2024).

Fard Ahmad Pourya, Hosseini Mohammad Mahdi, Sweeny Todd D. 2024. "AffectNet+: A Database for Enhancing Facial Expression Recognition with Soft-Labels." Retrieved from https://arxiv.org/pdf/2410.22506 (accessed December 7, 2024).

Fontes Luiz Maurilio Aires Moura Silva. 2023. "Scaleddown EfficientNet for Facial Expression Recognition." Retrieved from https://core.ac.uk/download/pdf/599045898.pdf (accessed December 7, 2024).

Gandhi Aneri, et al. 2023. "Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges, and Future Directions." Retrieved from http://w.sentic.net/multimodal-sentiment-analysis-review.pdf (accessed December 7, 2024).

Gu Alan. 2024. "A Facial and Vocal Emotion Recognition System Based on Deep Learning." Applied and Computational Engineering. Retrieved from https://www.ewadirect.com/proceedings/ace/article/view/15740 (accessed December 7, 2024).

Gupta Shikhar, Kumar Pramod, Tekchandani Rajkumar Kumar. 2023. "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models." Multimedia Tools and Applications. Retrieved from https://link.springer.com/article/10.1007/s11042-022-13558-9 (accessed December 7, 2024).

Hashky Ahmed, et al. 2024. "Multi-Modal User Modeling for Task Guidance: A Dataset for Real-Time Assistance with Stress and Interruption Dynamics." Retrieved from https://ieeexplore.ieee.org/ (accessed December 7, 2024).

Imani Mohammad, Montazer Gholam Ali. 2019. "A survey of emotion recognition methods with emphasis on E-Learning environments." Journal of Network and Computer Applications. Retrieved from https://www.sciencedirect.com/science/article/pii/S1084804519302759 (accessed December 7, 2024).

Kim Yunki, Soyata Tolga, Behnagh Razieh. 2018. "Towards Emotionally Aware AI Smart Classroom: Current Issues and Directions for Engineering and Education." Retrieved from https://ieeexplore.ieee.org/document/8253436 (accessed December 7, 2024).

Luckin Rose, Holmes Wayne. 2016. Intelligence Unleashed: An Argument for AI in Education. Pearson Education. Retrieved from https://discovery.ucl.ac.uk/id/eprint/1475756/ (accessed December 7, 2024).

Maoz Haim. 2024. "Transforming Education: Integrating AI-Driven Adaptation and Multimodal Approaches for Advanced Engineering Skills." Proceedings of the 2024 IEOM Tokyo Conference. Retrieved from https://ieomsociety.org/proceedings/2024tokyo/84.pdf (accessed December 7, 2024).

Nye Benjamin D., Graesser Arthur C., Hu Xiangen. 2014. "AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring." International Journal of Artificial Intelligence in Education. Retrieved from https://link.springer.com/content/pdf/10.1007/s40593-014-0029-5.pdf (accessed December 7, 2024).

Pappa Dimitra, et al. 2024. "Technology-Enhanced Education in Schools: Challenges to Well-Being." Retrieved from https://library.iated.org/view/PAPPA2024TEC (accessed December 7, 2024).

Pedro Francesc, et al. 2019. "Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development." Retrieved from http://repositorio.minedu.gob.pe/handle/20.500.12799/6533 (accessed December 7, 2024).

Picard Rosalind W. 1997. Affective Computing. MIT Press. Retrieved from https://affective.media.mit.edu/publications.php (accessed December 7, 2024).

Poria Soujanya, Cambria Erik, Hussain Amir, Huang Guang-Bin. 2015. "Towards an Intelligent Framework for Multimodal Affective Data Analysis." Neural Networks 63: 68-81. Retrieved from https://dspace.stir.ac.uk/bitstream/1893/21310/1/Neural%20Networks%202014.pdf (accessed December 7, 2024).

Rawat Rahul, Soundararajan Sriram. 2024. "A Comprehensive Development of Human Emotion Detection based on Facial Features using Digital Image Processing Methodology." Retrieved from https://ieeexplore.ieee.org/abstract/document/10511123/ (accessed December 7, 2024).

Sharma Kshitij, Papamitsiou Zacharoula. 2019. "Building Pipelines for Educational Data Using AI and Multimodal Analytics: A Grey-Box Approach." British Journal of Educational Technology. Retrieved from https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.12854 (accessed December 7, 2024).

Utami Puji, Hartanto Rudy, Soesanti Indah. 2022. "The EfficientNet Performance for Facial Expressions Recognition." Retrieved from https://ieeexplore.ieee.org/abstract/document/10053007/ (accessed December 7, 2024).

Waldner Daniel, Mitra Siddhartha. 2024. "Pairwise Discernment of AffectNet Expressions with ArcFace." Retrieved from https://arxiv.org/pdf/2412.01860 (accessed December 7, 2024).

Yu Shengnan, Androsov Artem, Yan Han, Chen Yuan. 2024. "Bridging computer and education sciences: A systematic review of automated emotion recognition in online learning environments." Computers & Education. Retrieved from https://www.sciencedirect.com/science/article/pii/S0360131524001258 (accessed December 7, 2024).

Zuo Enrui, Wu Wei. 2024. "Multi-Physiological Signal Fusion for Objective Emotion Recognition in Educational Human-Computer Interaction." Frontiers in Public Health. Retrieved from https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2024.1492375/full (accessed December 7, 2024).