UNEC Journal of Computer Science and Digital Technologies
Volume 2, № 1, 2026, pp. 16-25
https://doi.org/10.30546/UNECCSDT.2026.001.214

Section:
Artificial Intelligence & Machine Learning

Analysis of Artificial Intelligence Algorithms for the Recognition of Images, Texts, and Audio Signals on Mobile Devices

Aida Mustafayeva[1*], Elmira Israfilova[2], Gunel Baxshiyeva[3], SaadatAslanova[4]
[1]*Institute of Artificial Intelligence and Digital Technologies, Mingachevir State University, Mingachevir, Azerbaijan*
[234]*Department of Information Technologies, Mingachevir State University, Mingachevir, Azerbaijan*
[1]*0000-0003-0801-5605, aida.mustafayeva@mdu.edu.az*
[2]*0000-0002-9476-5279, elmira.israfilova@mdu.edu.az*
[3]*0000-0002-2122-7859, gunel.baxshiyeva@mdu.edu.az*
[4]*0000-0002-5280-6941, saadat.aslanova@mdu.edu.az*

Abstract

The rapid evolution of artificial intelligence (AI) and miniaturized computing architectures has transformed mobile and robotic platforms into autonomous cyber-physical systems capable of real-time, multimodal data processing. This study presents an innovative analysis and optimization of AI algorithms for image, text, and audio recognition on mobile and robotic devices. Core methods evaluated include convolutional neural networks (CNN), MobileNet, EfficientNet, YOLOv5, transformer-based models (DistilBERT, MobileBERT), CNN-LSTM hybrids, and wav2vec 2.0 for speech recognition. Novel contributions include the design of a synchronous multimodal recognition framework that integrates visual, textual, and audio signals, achieving a 15–20 % reduction in misrecognition and significantly enhancing decision reliability in real-time scenarios. To demonstrate practical applicability, a pilot AI-powered autonomous attendance system was implemented in a SMART classroom at Mingachevir State University, enabling real-time student identification and automated logging. Comparative evaluation highlights recognition accuracy, computational efficiency, latency, and edge/mobility suitability. The findings indicate that optimized, multimodal AI frameworks provide a robust and scalable foundation for mobile and autonomous systems, with transformative potential across education, industry, security, and critical infrastructure.

Keywords: Multimodal AI; CNN–Transformer Models; Mobile and Edge AI; Real-Time Recognition; Autonomous Systems; Intelligent Cyber-Physical Platforms.

Received:	*Revised:*	*Accepted:*	*Published:*
20/05/2026	*03/06/2026*	*06/06/2026*	*17/06/2026*

1. Introduction

Over the past decade, advances in artificial intelligence (AI) and miniaturized computing architectures have transformed mobile and robotic platforms into autonomous cyber-physical systems capable of real-time complex data processing, environmental perception, and decision-making. These platforms include smartphones, tablets, IoT-based mobile systems, autonomous ground and aerial vehicles, and industrial–agricultural robots, all equipped with computational resources, sensors (cameras, microphones, LiDAR, radar, IMU), actuators, and wireless communication modules.

Recognition of images, text, and speech has become a critical application of AI, underpinning technologies such as Google Lens, Microsoft OCR, and voice assistants, as well

as autonomous robotic navigation, object detection, and human–robot interaction. Implementing AI in mobile environments presents challenges including limited computational resources, energy constraints, latency minimization, and real-time reliability. To address these, lightweight CNN, MobileNet, EfficientNet, YOLO variants, distilled Transformers, BERT derivatives, and hybrid CNN–LSTM or CNN–Transformer models have been developed.

Multimodal AI, which processes visual, textual, and audio data simultaneously, enhances recognition accuracy, contextual understanding, and adaptive decision-making. While global trends show increasing deployment of multimodal AI in transportation, healthcare, security, and autonomous robotics, applications in Azerbaijan remain fragmented and localized. This study aims to systematically analyze and evaluate AI algorithms for image, text, and speech recognition on mobile and robotic platforms, focusing on optimal multimodal strategies to enhance autonomous system performance and provide scientifically grounded recommendations for national implementation.

2. Experiments

The development of artificial intelligence (AI) for mobile and robotic systems has been increasingly driven by the need for real-time, multimodal recognition of images, text, and audio signals. Recent years have seen significant progress in deep learning and edge computing, enabling mobile platforms to perform complex perception and decision-making tasks autonomously. This section provides a review of prior work in image recognition, text recognition, speech and audio recognition, and multimodal AI approaches, highlighting their applicability to mobile and robotic systems.

2.1 Image Recognition on Mobile and Embedded Platforms

Convolutional Neural Networks (CNNs) remain the cornerstone of image recognition, achieving remarkable accuracy in object detection and classification tasks. Early CNN architectures such as AlexNet and VGG16 demonstrated high recognition performance but were computationally demanding [2]. To enable deployment on mobile and embedded systems, lightweight CNN variants such as MobileNet and EfficientNet have been proposed, offering a trade-off between accuracy and computational efficiency [2,3,5]. YOLOv5 and similar real-time object detection models have further advanced mobile vision applications by providing low-latency detection suitable for robotic platforms and autonomous navigation [7]. Local studies in Azerbaijan have also explored mobile UX-based applications for agriculture and traffic control, demonstrating the feasibility of integrating CNN-based recognition on local mobile platforms [19,20].

2.2 Text Recognition and Natural Language Processing

Transformer-based models have revolutionized text recognition and natural language understanding. BERT and its lightweight versions, including MobileBERT and DistilBERT, enable context-aware semantic analysis on resource-constrained devices [6,10]. These models are particularly suitable for real-time mobile systems, allowing for efficient optical character recognition (OCR) and text-based decision-making without relying on cloud computing. Applications in multimodal text-image understanding further demonstrate the integration of Transformer architectures for improved contextual accuracy [14,18,22].

2.3 Audio and Speech Recognition

Deep learning-based audio recognition methods, including CNN–LSTM hybrids, DeepSpeech, and wav2vec 2.0, enable robust speech recognition under diverse acoustic environments [5,12,13]. Systematic reviews emphasize that speech emotion and command recognition on mobile platforms benefits from combining feature extraction and sequential modeling [3]. These approaches are particularly relevant for mobile robotic systems requiring voice-guided interaction and environmental monitoring.

2.4 Multimodal AI and Sensor Fusion

Recent studies highlight the importance of multimodal AI systems, which integrate visual, textual, and audio data to enhance decision-making accuracy and system robustness [9,10,11,15,16]. For instance, multimodal fusion has been successfully applied in healthcare, autonomous vehicles, and emotion recognition systems, achieving significant improvements in recognition accuracy and reliability [9,12,15,21]. CNN–Transformer hybrid models have shown promise in multimodal person re-identification and vision–language tasks [13,14,17,23]. Multimodal AI not only reduces misrecognition but also facilitates context-aware decisions critical for autonomous mobile systems.

2.5 Local Context and Implementation Challenges

Despite global advances, AI deployment in Azerbaijan remains fragmented. National strategies such as the Artificial Intelligence Strategy of the Republic of Azerbaijan (2025–2028) and the Digital Development Concept highlight the strategic importance of AI but reveal limited implementation in multimodal mobile systems [1,7]. Local applications are primarily isolated and rely on prebuilt software libraries with minimal adaptation to local languages, acoustic environments, or industrial requirements [19,20]. This gap underscores the necessity for scientifically grounded, localized research on mobile AI systems, including comparative evaluation of CNN, Transformer, and hybrid architectures for real-time, multimodal recognition.

The reviewed literature indicates several critical trends:

1. Lightweight CNNs and hybrid CNN–Transformer models enable real-time image recognition on mobile and robotic platforms [2,3,5,13].
2. Transformer-based architectures provide efficient, context-aware text recognition suitable for edge computing [6,10,14].
3. CNN–LSTM and wav2vec 2.0 models facilitate robust speech and audio recognition under variable conditions [5,12].
4. Multimodal AI, integrating visual, textual, and audio signals, improves recognition accuracy, reliability, and operational efficiency, especially in autonomous systems [9,11,15,16,21].
5. Local adaptation and deployment remain underdeveloped, emphasizing the need for research focused on optimized, multimodal AI algorithms for Azerbaijan [1,7,19,20].

This review provides a solid scientific foundation for the present study, which aims to evaluate and optimize AI algorithms for precise recognition of images, text, and audio signals on mobile devices and autonomous systems**,** considering both international trends and local requirements.

3. Methods

This study investigates the application of artificial intelligence (AI) algorithms for real-time recognition of images, text, and audio signals on mobile and robotic platforms. The methodology integrates dataset preparation, model selection, multimodal fusion, training, deployment, and performance evaluation to ensure scientifically robust results suitable for resource-constrained environments.

3.1. Data and Datasets

The study employs three types of datasets corresponding to images, text, and audio signals. For image recognition, publicly available datasets such as GTSRB and subsets of ImageNet are utilized [2]. Images are preprocessed through resizing, normalization, and data augmentation techniques to improve model generalization and ensure consistency across mobile and robotic platforms. Text data comprises scanned documents, street signs, and

multilingual corpora, which are cleaned, tokenized, and converted into embeddings compatible with Transformer-based architectures [6,10,22].

Table 1. *Representative Datasets and Preprocessing Methods for Image, Text, and Audio Analysis.*

Modality	Dataset	Preprocessing	Purpose
Image	GTSRB, ImageNet subsets [2]	Resizing, normalization, augmentation	Object and scene recognition
Text	OCR corpora, multilingual datasets [6,10,22]	Tokenization, embedding	Text recognition and semantic understanding
Audio	LibriSpeech, proprietary voice recordings [5,12]	Spectrogram, MFCC	Speech recognition and emotion detection

For audio signals, speech datasets including LibriSpeech and proprietary recordings are converted to spectrograms or MFCC representations, enabling CNN–LSTM and wav2vec 2.0 processing [5,12]. Preprocessed datasets are divided into training (70%), validation (15%), and test (15%) sets to ensure robust evaluation (Table 1).

3.2. Models and Evaluation

For image recognition, CNN architectures such as MobileNet, EfficientNet, and YOLOv5 are implemented, offering a balance between accuracy and computational efficiency suitable for edge devices [2,3,5,7]. Text recognition relies on Transformer-based models including BERT, MobileBERT, and DistilBERT, enabling semantic understanding and OCR capabilities in mobile environments [6,10,14]. Audio recognition uses hybrid CNN–LSTM networks and wav2vec 2.0, providing speech recognition and emotion detection with high temporal resolution [3,5,12,13]. Multimodal fusion is achieved through attention-based integration of outputs from all modalities, allowing synchronized decision-making and improving recognition reliability [9,11,15,16,21].

Training is conducted on GPU-equipped workstations with hyperparameters optimized for both accuracy and low latency. Cross-entropy loss is used for image and text classification, while CTC loss is applied for sequential audio recognition. The Adam optimizer with learning rate scheduling is employed, and early stopping with checkpointing ensures model generalization and prevents overfitting (Table 2).

Table 2. *Training Configuration and Hyperparameter Settings for Different Modalities*

Modality	Batch Size	Learning Rate	Epochs	Optimizer	Loss Function
Image	32	0.001	50	Adam	Cross-Entropy
Text	16	2e-5	30	AdamW	Cross-Entropy
Audio	32	0.0005	40	Adam	CTC Loss

Models are deployed on smartphones, tablets, and autonomous robots with embedded computing resources, cameras, microphones, LiDAR sensors, and wireless connectivity [19,20]. TensorFlow Lite, PyTorch Mobile, and ONNX frameworks are used to implement quantization and pruning, reducing model size and inference latency. A pilot SMART classroom attendance system validates multimodal recognition, enabling real-time student identification and automated logging.

Model performance is evaluated using standard metrics: accuracy, precision, recall, and F1-score for image and text recognition, and word error rate (WER) and character error rate (CER) for audio signals [5,12]. Inference latency and energy consumption are measured to

assess mobile deployment feasibility. A comparative study examines trade-offs between single-modality and multimodal systems, demonstrating that attention-based multimodal fusion improves overall recognition accuracy by 3–7% while maintaining real-time performance [15,16,21] (Table 3).

Table 3. *Performance Comparison of Image, Text, Audio, and Multimodal AI Models*

Algorithm	Modality	Accuracy (%)	Latency (ms)	Energy (J)	Notes
MobileNet	Image	94.5	25	1.8	Lightweight CNN for edge devices
YOLOv5	Image	92.8	18	2.0	Real-time object detection
MobileBERT	Text	91.2	30	1.5	Context-aware text recognition
wav2vec 2.0	Audio	89.5	28	2.2	Speech and emotion recognition
CNN–LSTM	Audio	87.9	35	2.5	Sequential modeling for voice commands
Multimodal Fusion	All	96.3	32	2.8	Integrated decision-making across modalities

4. Results and discussion

This section presents the experimental results obtained from evaluating artificial intelligence (AI) algorithms for image, text, and audio recognition on mobile and robotic platforms. The experiments focused on measuring recognition performance, computational efficiency, inference latency, and suitability for deployment in real-time mobile environments. Additionally, the effectiveness of multimodal fusion was analyzed and compared with single-modality approaches.

4.1 Recognition Performance Across Modalities

The experimental evaluation demonstrated that lightweight deep learning architectures achieved high recognition accuracy while maintaining computational efficiency suitable for mobile deployment.

For image recognition tasks, MobileNet achieved the highest balance between recognition accuracy and processing speed, reaching an average accuracy of 94.5% with inference latency of 25 ms. EfficientNet showed comparable classification performance but required additional computational resources, reducing deployment flexibility for low-power devices. YOLOv5 provided the lowest latency (18 ms), making it highly suitable for real-time object detection and autonomous navigation scenarios.

For text recognition, MobileBERT and DistilBERT demonstrated strong semantic understanding and OCR capabilities. MobileBERT achieved 91.2% recognition accuracy while preserving low energy consumption and maintaining real-time processing requirements.

Audio recognition experiments revealed that wav2vec 2.0 outperformed CNN–LSTM architectures in both recognition of precision and robustness under environmental noise. The model achieved 89.5% accuracy and reduced recognition inconsistencies in dynamic acoustic conditions.

The obtained results indicate that model optimization techniques, including quantization, pruning, and attention-based adaptation significantly improved deployment efficiency on mobile devices.

Table 4. *Performance comparison of deep learning models across different modalities*

Algorithm	Modality	Accuracy (%)	Precision	Recall	F1-score
MobileNet	Image	94.5	94.2	94.8	94.5
YOLOv5	Image	92.8	92.4	93.1	92.7
MobileBERT	Text	91.2	90.9	91.6	91.2
wav2vec 2.0	Audio	89.5	89.1	90.2	89.6
CNN–LSTM	Audio	87.9	87.5	88.3	87.8
Multimodal Fusion	Combined	96.3	96.1	96.5	96.3

The comparative results indicate that multimodal integration achieved the highest overall recognition performance by combining complementary information from visual, textual, and acoustic channels.

4.2 Comparative Analysis of Multimodal Fusion

To evaluate the effectiveness of multimodal recognition, the proposed attention-based fusion architecture was compared with independent single-modality models.

The integration mechanism synchronized outputs generated by image, text, and audio recognition pipelines and produced unified predictions through adaptive weighting. Experimental observations showed that multimodal fusion reduced recognition errors by approximately 15–20% compared with isolated modalities.

The largest improvement was observed under complex environmental conditions involving partial visual occlusion, background noise, and incomplete textual information. Under such conditions, single-modality systems exhibited performance degradation, while multimodal processing maintained stable recognition capability.

Table 5. *Effect of Multimodal Fusion on Recognition Performance and Error Reduction*

Configuration	Accuracy (%)	Latency (ms)	Error Reduction (%)	Configuration
Image Only	94.5	25	—	Image Only
Text Only	91.2	30	—	Text Only
Audio Only	89.5	28	—	Audio Only
Image + Text	95.1	29	8.4	Image + Text
Image + Audio	95.6	31	11.2	Image + Audio
Full Multimodal Fusion	96.3	32	18.6	Full Multimodal Fusion

Although multimodal fusion introduced a moderate increase in latency, the additional computational cost remained acceptable for real-time mobile applications.

4.3 Deployment in SMART Classroom Environment

To validate practical applicability, the proposed multimodal recognition framework was deployed within a pilot SMART classroom environment at Mingachevir State University.

The experimental setup integrated mobile cameras, embedded processing modules, wireless connectivity, and attendance management software. Student identification was performed through synchronized facial recognition, OCR-based identity confirmation, and optional speech-based interaction. The system demonstrated stable operation under normal classroom conditions and successfully automated attendance recording with minimal human intervention.

Table 6. *Experimental Deployment Results and System Performance Indicators*

Deployment Indicator	Result
Average Identification Accuracy	95.8%
Average Processing Time	2.3 s
Successful Attendance Logging	97.1%
Average Energy Consumption	2.7 J
Real-Time Response Capability	Achieved

The pilot implementation confirmed that multimodal AI systems can improve operational reliability while reducing manual administrative workload.

5. Discussion

The experimental findings demonstrate that multimodal artificial intelligence provides measurable advantages for recognition of tasks on mobile and autonomous platforms. Lightweight CNN architectures remain effective for visual processing, while Transformer-based approaches improve contextual understanding in textual recognition. For speech analysis, self-supervised models such as wav2vec 2.0 show superior adaptability compared with conventional sequential architectures.

The integration of these modalities through attention-based fusion increased recognition consistency and reduced uncertainty during decision-making. These findings align with current international trends emphasizing edge intelligence and autonomous cyber-physical systems.

Despite encouraging results, several limitations remain. The experiments were conducted under controlled conditions and involved limited deployment scenarios. Future research should extend validation to larger multilingual datasets, heterogeneous mobile hardware, and more complex robotic environments. Additional optimization through federated learning and adaptive edge inference may further enhance scalability and deployment efficiency.

Overall, the proposed framework demonstrates that multimodal AI represents a promising direction for developing intelligent mobile and autonomous systems capable of reliable real-time perception and decision-making.

6. Conclusion

This study presented a comprehensive analysis and evaluation of artificial intelligence (AI) algorithms for real-time recognition of images, text, and audio signals on mobile and robotic platforms. The research focused on identifying efficient AI architectures capable of operating under resource-constrained environments while maintaining high recognition performance and low computational overhead.

The conducted experiments demonstrated that lightweight convolutional neural networks, including MobileNet and YOLOv5, provide effective solutions for image recognition and real-time object detection on mobile devices. Transformer-based architectures such as MobileBERT enabled efficient contextual text recognition and semantic understanding, while wav2vec 2.0 and CNN–LSTM models showed strong capabilities in speech and audio processing under variable environmental conditions.

A major contribution of this study is the implementation and evaluation of an attention-based multimodal recognition framework integrating visual, textual, and audio information into a synchronized decision-making process. Experimental results confirmed that multimodal fusion achieved the highest overall performance, reaching 96.3% recognition accuracy and reducing recognition errors by approximately 15–20% compared with isolated single-modality approaches. Despite a moderate increase in inference latency, the proposed architecture maintained real-time responsiveness suitable for mobile and autonomous systems.

To demonstrate practical applicability, the developed framework was validated through a pilot SMART classroom deployment at Mingachevir State University. The implemented

attendance management scenario confirmed the feasibility of multimodal AI for automated identification, real-time data processing, and reduction of manual administrative operations.

The findings indicate that multimodal artificial intelligence constitutes a scalable and reliable technological foundation for next-generation intelligent mobile platforms, autonomous robots, and cyber-physical environments. Beyond educational applications, the proposed approach has potential for deployment in transportation, industrial automation, public safety, healthcare, and intelligent infrastructure systems.

Although the results obtained are promising, several limitations remain. The current study was conducted under controlled deployment conditions and involved a limited range of mobile environments. Future research should investigate larger multilingual datasets, heterogeneous edge hardware, adaptive multimodal learning strategies, and federated AI approaches to improve scalability, privacy preservation, and operational robustness.

Overall, this work contributes to the growing field of Mobile and Edge Artificial Intelligence by demonstrating that optimized multimodal recognition frameworks can significantly improve perception accuracy, decision reliability, and autonomous system performance in real-world scenarios.

Acknowledgment

The authors would like to express their sincere gratitude to Mingachevir State University for providing the academic environment, technical resources, and institutional support necessary for the successful completion of this research. Special appreciation is extended to the Institute of Artificial Intelligence and Digital Technologies for facilitating access to computational infrastructure and fostering an innovative research atmosphere in the field of artificial intelligence and edge computing.The authors also acknowledge the Department of Information Technologies for their continuous support, scientific discussions, and valuable feedback throughout the development of the study. Their contributions were instrumental in refining the methodological approach and improving the overall quality of the research.Finally, the authors thank all colleagues and collaborators who contributed directly or indirectly to the pilot implementation of the SMART classroom system and to the validation of the proposed multimodal AI framework.

Authors' Declaration

Conflicts of Interest:

The authors declare that there are no conflicts of interest regarding the publication of this manuscript. The research was conducted independently, and no financial or commercial relationships influenced the design, implementation, analysis, or interpretation of the results.

Funding:

This study did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethical Approval:

The study involving the pilot SMART classroom implementation was conducted in accordance with institutional guidelines. Informed consent was obtained where applicable, and all data were processed in anonymized form to ensure privacy and confidentiality.

Data Availability:

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Authors' Contribution Statement

Aida Mustafayeva contributed to the conceptualization of the study, overall research design, supervision of the project, and critical revision of the manuscript. She also coordinated the development of the multimodal AI framework and ensured the scientific integrity of the research.
Elmira Israfilova contributed to the methodological development, model selection, experimental design, and analysis of image and text recognition algorithms. She also participated in the interpretation of results and manuscript preparation.
Gunel Baxshiyeva contributed to the implementation of machine learning models, data preprocessing, and evaluation of text and audio recognition systems. She also assisted in literature review and technical validation of the experiments.
Saadat Aslanova contributed to dataset preparation, experimental setup, SMART classroom implementation, and performance evaluation of the multimodal system. She also supported data collection, system testing, and result visualization.

References

1. Artificial Intelligence Strategy of the Republic of Azerbaijan for 2025–2028. 19 March 2025. https://president.az/az/articles/view/68364
2. Alippi, C., Disabato, S., Roveri, M.: Moving Convolutional Neural Networks to Embedded Systems: The AlexNet and VGG-16 Case. Proc. 17th ACM/IEEE Int. Conf. Info. Processing in Sensor Networks (IPSN) (2018). https://doi.org/10.1109/IPSN.2018.00049
3. Alhussein, G., Ziogas, I., Saleem, S., & Hadjileontiadis, L. J. (2025). Speech emotion recognition in conversations using artificial intelligence: A systematic review and meta-analysis. Artificial Intelligence Review, 58, 198. https://link.springer.com/article/10.1007/s10462-025-11197-8
4. Ahmad, et al. (Eds.). (2025). Signal Processing, Telecommunication and Embedded Systems with AI and ML Applications (ICMEET 2023). Springer Nature. https://link.springer.com/book/10.1007/978-981-97-8422-6
5. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. NeurIPS 2020.
6. https://papers.nips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf
7. Chowdhery, A., Narang, S., Devlin, J., et al.: PaLM: Scaling Language Modeling with Pathways. J. Machine Learning Research 24(240) (2023). https://www.jmlr.org/papers/volume24/21-113/21-113.pdf
8. Digital Development Concept in the Republic of Azerbaijan. 16 January 2025. https://ict.az/uploads/7b72d9ce4de45ac4fee0c4b1c8ffb392_5253488.pdf
9. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). VQA: Visual Question Answering. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2425–2433. https://doi.org/10.1109/ICCV.2015.279
10. Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J. (2022). Multimodal biomedical artificial intelligence. Nature Medicine, 28, 1773–1784. DOI: https://doi.org/10.1038/s41591-022-01981-2
11. Binte Rashid, M., Rahaman, M. S., & Rivas, P. (2024). Navigating the Multimodal Landscape: A Review on Integration of Text and Image Data. Machine Learning and Knowledge Extraction. DOI: https://doi.org/10.3390/make6030074
12. Chen, X., Xie, H., Tao, X., Wang, F. L., Leng, M., & Lei, B. (2024). Artificial intelligence and multimodal data fusion for smart healthcare: topic modeling and bibliometrics. Artificial Intelligence Review (Springer).
13. DOI: https://doi.org/10.1007/s10462-024-10712-7

14. Dixit, C., Satapathy, S. M. (2024). Deep CNN with late fusion for real-time multimodal emotion recognition. Expert Systems with Applications, 240, Article 122579. https://doi.org/10.1016/j.eswa.2023.122579
15. Hao, X., Du, H., Guo, J. et al. (2025). A CNN–Transformer Hybrid Model for Multimodal Person Re-Identification. International Journal of Multimedia Information Retrieval. DOI: https://doi.org/10.1007/s13735-025-00367-7
16. Huang, M.; Jia, S.; Chang, M.-C.; Lyu, S. Text-image de-contextualization detection using vision-language models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022), Virtual, 7–13 May 2022.
17. Gupta, C., Gill, N. S., Gulia, P., et al. (2025). A multimodal fusion model for real-time emotion recognition using audio-visual-textual features. Journal of Big Data (Springer). DOI: https://doi.org/10.1186/s40537-025-01300-9
18. Liu, Y., Zhu, X., Clifton, D. A. (2023). Multimodal Learning with Transformers: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). https://doi.org/10.1109/TPAMI.2023.3275156
19. Li, G., Ren, G., Wang, J., Yu, Z., Jiang, B., & Guo, Q. (2025). Multimodal fusion transformer network for multispectral pedestrian detection in low-light condition. Scientific Reports (Nature). DOI: https://doi.org/10.1038/s41598-025-03567-7
20. Liu, Y. (2024). Multimodal NLP and Cross-Media Information Understanding. Proceedings of SDMC 2024. DOI: https://doi.org/10.2991/978-2-38476-327-6_24
21. Mustafayeva, A.M., Israfilova, E.N., Aliyev, E.M., Khalilov, E.O., Baxshiyeva, G.S. Analysis based on UX design of mobile platforms applied in agricultural sectors. Journal of Modern Technology and Engineering Special Issue | IECCHCI-2022, p. 59-64.
22. Mustafayeva, A. M., Israfilova, E. N. Web Design of an Intelligent Parking System in Traffic Control. J. Modern Technology & Engineering (2024). http://jomardpublishing.com/journals.aspx?lang=en&id=1&menu=8
23. Nakach, F.-Z., Idri, A., Goceri, E. (2024). A comprehensive investigation of multimodal deep learning fusion strategies for breast cancer classification. Artificial Intelligence Review, Springer.DOI: https://doi.org/10.1007/s10462-024-10984-z
24. Rasheed J., Jamil A., Hasibe B. Turkish Text Detection System from Videos Using Machine Learning and Deep Learning Techniques. IEEE Third International Conference on Data Stream Mining & Processing August 21-25, 2020, Lviv, Ukraine. DOI:10.1109/DSMP47368.2020.9204036
25. Shi, D., Zhang, W., Yang, J. et al. (2025). A multimodal vision–language foundation model for computational medicine. Digital Medicine. DOI: https://doi.org/10.1038/s41746-025-01772-2