DOI: 10.20998/2079-0023.2025.02.12
UDC 004.62

O. V. ZHEREBETSKYI, PhD student at the Department of Artificial Intelligence Systems, Lviv Polytechnic National University, Lviv, Ukraine; e-mail: oleh.v.zherebetskyi@lpnu.ua; ORCID: https://orcid.org / 0009-0004-6259-7065
O. A. BASYSTIUK, Candidate of Technical Sciences (PhD), Senior Lecturer at the Department of Artificial Intelligence Systems, Lviv Polytechnic National University, Lviv, Ukraine; e-mail: oleh.a.basystiuk@lpnu.ua; ORCID: https://orcid.org /0000-0003-0064-6584

INTEGRATION OF HETEROGENEOUS DATA USING ARTIFICIAL INTELLIGENCE METHODS

Modern AI development and multimodal data analysis methods are gaining critical importance due to their ability to integrate information from diverse sources, including text, audio, sensor signals, and images. Such integration enables systems to form a richer and more context-aware understanding of complex environments, which is essential for domains such as healthcare diagnostics, adaptive education technologies, intelligent security systems, autonomous robotics, and various forms of human-computer interaction. Multimodal approaches also enable AI models to compensate for the limitations inherent in individual modalities, thereby enhancing robustness and resilience to noise or incomplete data. The study employs theoretical analysis of scientific literature, comparative classification of multimodal architectures, systematization of fusion techniques, and formal generalization of model design principles. Additionally, attention is given to evaluating emerging paradigms powered by large-scale foundation models and transformer-based architectures. The primary methods and models for processing multimodal data are summarized, covering both classical and state-of-the-art approaches. Architectures of early (feature-level), late (decision-level), and hybrid (intermediate) fusion are described and compared in terms of flexibility, computational complexity, interpretability, and accuracy. Emerging solutions based on large multimodal transformer models, contrastive learning, and unified embedding spaces are also analyzed. Special attention is paid to cross-modal attention mechanisms that enable dynamic weighting of modalities depending on task context. The study determines that multimodal systems achieve significantly higher accuracy, stability, and semantic coherence in classification, detection, and interpretation tasks when modalities are properly synchronized and fused using adaptive strategies. These findings underscore the promise of further research toward scalable architectures capable of real-time multimodal reasoning, improved cross-modal transfer, and context-aware attention mechanisms.

Keywords: multimodality, artificial intelligence, emotion classification, fusion architectures, audio-video-text processing, transformers, cross-modal attention.

Introduction. In today's IT environment, there has been a sharp increase in the volume of different types of data–text messages, audio, and video streams–coming from web services, sensors, and social media. Multimodal approaches, inspired by the human ability to perceive different channels of information simultaneously, allow us to build models with a deeper understanding of context [1]. The fusion of information from multiple modalities. It enables the creation of more robust and informative systems: in particular, recent studies have demonstrated that multimodal models significantly outperform single-channel approaches in various tasks, ranging from question answering to medical diagnosis [2]. For example, in the field of cybersecurity and information reliability, it has been demonstrated that fake news often incorporates combined media elements to manipulate readers' perceptions [3]. This fact underscores the need for tools that can simultaneously analyze text descriptions and accompanying visual/audio materials.

The availability of heterogeneous multimodal data plays a key role in the development of IT and AI. On the one hand, modern machine learning architectures can flexibly process different data formats, and in theory, this opens up new opportunities for intelligent applications. On the other hand, this approach enables artificial intelligence systems to approximate the human way of perceiving reality–a person simultaneously analyzes visual images, sound signals, and verbal information. As the researchers emphasize, integrating information from multiple modalities is sometimes the only way to solve the problem of object recognition or semantic interpretation fully.

As researchers point out, integrating information from multiple modalities is sometimes the only way to fully solve the task of object recognition or semantic interpretation of scenes [4, 5]. For example, when detecting false information, matching the text content with the image is critically important – discrepancies between modalities alone can be a sign of manipulation [6]. Thus, the processing and combination of text, sound, and images are integral parts of modern AI research, significantly improving the quality of analytical and diagnostic systems.

Current challenges and trends. Recent global events have significantly increased the need for multimodal solutions. First, the COVID-19 pandemic has accelerated the transition to remote work and learning, with video conferencing and online platforms becoming the primary channels of communication. Interactive environments require systems that can simultaneously process video, audio, and text streams. As observers note, interaction in distance learning is inextricably linked to multimodal interfaces. Secondly, large-scale crises are accompanied by an avalanche of information from social networks and the media. In such conditions, disinformation is often spread using synchronized multimedia content [7, 8]. On the other hand, the development of autonomous systems–from driverless cars to robots–involves combining different types of sensor data (such as cameras, LiDAR, radar, and microphones) to achieve a comprehensive understanding of the environment. Reviews in the field of auto-recognition emphasize that a multimodal sensor fusion system–the merging of data from cameras, LiDAR, and radars–significantly increases the reliability of detecting moving objects [9].

Research Article: This article was published by the publishing house of ***NTU "KhPI"*** in the collection "Bulletin of the National Technical University "KhPI" Series: System analysis, management and information technologies." This article is distributed under a Creative Common Creative Common Attribution (CC BY 4.0). **Conflict of Interest:** The author/s declared no conflict of interest.

Traditional and modern multimodal processing methods. We should not forget about the rapid growth of multimedia content on social networks, where a combined analysis of images, audio, and text plays a crucial role, making multimodal technologies particularly in demand across all areas of information technology. Degree of research. Due to these requirements, numerous scientific reviews have been devoted to multimodal machine learning in recent years. For example, article [10] provides a thorough overview of modern methods of multimodal learning, their architectures, and key areas of application. In the field of biomedicine [11], there is a growing interest in combining visual images (such as CT and MRI) with clinical information to enhance diagnostic systems [12, 13]. New multimodal fusion algorithms are being intensively developed, particularly based on transformers with cross-attention mechanisms–they exhibit high accuracy but face scalability issues when combining more than two modalities. Recognition and classification tasks are being actively researched [14]: for example, detecting fake news using multimodal methods and recognizing emotions from facial images and speech intonations.

A review [15] shows a sharp increase in the number of publications in the field of multimodal disinformation since spring 2020. However, it also notes serious gaps, including the lack of a single agreed-upon terminology and methodology, as well as the absence of interdisciplinary research and international communities working at the intersection of computer science, linguistics, and political science. Technical problems include the synchronization of heterogeneous data and the high computational costs of deep learning algorithms. In general, it can be stated that the field of multimodal analysis is in a phase of active growth: some areas, such as visual-text models and large language-vision transformers, are currently in the spotlight, while others, such as the simultaneous analysis of more than three modalities and real-time processing of short videos, still require new solutions.

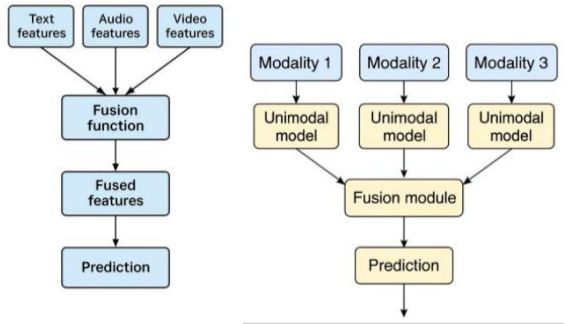

Fig. 1. Diagram of early, late fusion

For a clearer understanding of the principles of multimodal system construction, it is helpful to consider typical diagrams of the two main approaches to data fusion: early and late fusion. Above are structured illustrations of each option for integrating features.

Fig. 1 presents the key architectural differences between the approaches, including the point of feature merging, the level of interaction between modalities, and the location of decision-making. The choice of fusion strategy determines both the accuracy of the model and its resistance to the loss or distortion of individual modalities.

First, there is a fundamental difference in structure, scale, and temporal nature between different types of data: text, audio, and visual. This complicates the construction of a unified representation that would preserve the significant features of each modality without losing semantics. Formally, the process of combining modalities can be represented as a function:

$$h = f(x_{\text{text}}, x_{\text{audio}}, x_{\text{video}}) \quad (1)$$

where x_{text}, x_{audio}, x_{video} – are the feature vectors of the corresponding modalities, and f is the fusion function.

The choice of this function determines the system's architecture, but there is currently no universal approach suitable for all tasks.

Second, most modern models are limited to two modalities, while real data is often more complex [16]. Merging more than two sources of information leads to an exponential increase in computational costs, which creates significant technical difficulties when deploying such systems in practical conditions.

Third, the research community still lacks agreed-upon standards for selecting test sets, evaluation methods, and architectural design. This complicates the comparison of results and hinders progress in the development of generalized solutions [17]. The purpose of this article is to systematize scientific results related to methods of processing and integrating multimodal data using artificial intelligence. Such a review enables us to identify key architectural solutions, assess the effectiveness of fundamental approaches, and suggest directions for future research in this dynamic field.

The material is structured according to the principle of gradual detailing: first, the basic methods of representation and fusion of modalities are analyzed; then, modern software frameworks and application systems are considered; and finally, generalizations, limitations, and prospects for the development of a multimodal approach are formulated.

Main problems of multimodal systems:

- Noise and data interference. In real recordings, individual modalities can be significantly noisy. Background sounds, artifacts in images, and other factors complicate the accurate integration of data.
- Lack of synchronization. Different modalities often have their own time scales and frequencies, so their temporal alignment is non-trivial.
- Incomplete data. In many cases, some modalities are missing from the data, which worsens the results of typical models.
- Heterogeneity and incompatibility of modalities. Data from different sources have fundamentally different formats and dimensions, which requires special integration mechanisms.
- Scaling complexity. As the number of modalities increases, the complexity of the model and the amount of

necessary computations grow exponentially, complicating training and inference.

- Lack of open datasets. There is a shortage of high-quality multimodal datasets, particularly those containing authentic emotional and medical data. This limits the possibilities for researchers. As noted in the K-EmoPhone study, there is still a lack of open datasets collected in real-world conditions with labels for emotions and cognitive states.

Shortcomings of current approaches. Reviews and experimental results indicate several systemic shortcomings in current multimodal approaches. First, the large number of modalities complicates the construction of a generalized representation. As noted in the study [18], multimodal systems present unique challenges due to the heterogeneity of data sources and the interrelationships between modalities.

Criteria for comparison. This results in a significant increase in the number of model parameters and the training data requirements.

Second, most modern models are difficult to adapt to incomplete or missing modalities: in the absence of one of the channels, performance suffers significantly.

Third, the results of multimodal models are often difficult to interpret. As noted in article [19], the use of NLP and ML enables the extraction of additional information from different modalities. However, in real-world conditions, the task remains far from trivial, requiring significant data preprocessing, and the results should be interpreted with caution.

Finally, due to the high complexity of multimodal model systems, a vast number of training examples and computing resources are required. As the same researchers point out, further research is needed before these methods can be implemented at scale, indicating a dependence on large datasets and lengthy training.

Summary of disadvantages:

- Complexity of models in the presence of multiple modalities. Exponential growth of parameters.
- Vulnerability to missing or noisy data.
- Decreased accuracy with incomplete modalities.
- Low interpretability of results. Opacity of multimodal models.
- High demand for large training samples and computing resources.
- Complexity of coordinating and synchronizing heterogeneous data.
- Insufficiency of open multimodal datasets, especially with real-world scenarios.

A typical pipeline for multimodal emotion analysis:

1. Data collection. Create or utilize existing multimodal emotion datasets (audio, video, and text). For example, the RAVDESS dataset contains simultaneous audio and video recordings of actors expressing different emotions.

Data can be collected in a studio equipped with specialized equipment to ensure signal quality; for example, sound is recorded in a soundproof booth with background noise eliminated. Transcripts are usually obtained using automatic speech recognition (ASR) or manually.

2. Modality-specific preprocessing. At this stage, signals in each channel are cleaned up. Audio files are noise-cancelled, and inactive sections are cut out; the volume is then normalized.

Video frames are adjusted for lighting, face and/or gesture detection is performed, and irrelevant areas are cropped. The text is cleaned of punctuation, dialects, and redundant stop words, and then tokenized. For example, the RAVDESS dataset mentioned above was recorded in a professional studio to minimize noise.

3. Feature extraction. Numerical vectors are extracted from the prepared signals. For audio, these can include spectral features such as MFCC and energy coefficients in frequency bands.

For facial images, convolutional neural networks (CNNs) are commonly used: each frame is passed through a network (e.g., ResNet) and high-level output features are extracted. Body gestures are described by sets of joint coordinates, known as posture features. Text data is converted into word vectors: contextual embeddings (e.g., BERT/GPT) are used, or features are extracted using sequential models.

4. Alignment and synchronization. Since the temporal structure of features is different, they need to be aligned in terms of temporal context. Alignment methods are used, for example, such as joint framing of audio and video streams or aligning audio with text along lexical boundaries.

As an example, in their research, the authors divide the audio into segments based on the time of appearance of each word and linguistic boundaries, resulting in an aligned audio-text pair. This ensures that features from different modalities correspond to the same semantic segment.

5. Fusion module. After synchronization, features from all modalities are combined for further training. There are two types of fusion: early (feature-level) and late (decision-level). In early fusion, feature vectors are concatenated into a single, common vector; however, direct concatenation may not account for differences in size and framing.

In late fusion, each channel is processed separately, and then its predictions are combined. There are also hybrid architectures, such as partial fusion at an intermediate stage in deep networks.

6. Classifier. The fused features are fed into a classifier, such as a multilayer perceptron with a softmax shift at the output.

In training, the loss function is minimized, for example, by reducing the cross-entropy between the predicted and actual labels. Sometimes recurrent networks (LSTM) or SVM/Random Forest are used for final classification, depending on the approach.

7. Output. The final result is an emotion prediction. This is usually either a categorical label (e.g., “happiness,” “sadness,” “anger”) or a probability distribution across several emotional categories (softmax state).

The final label is selected as the one with the maximum probability.

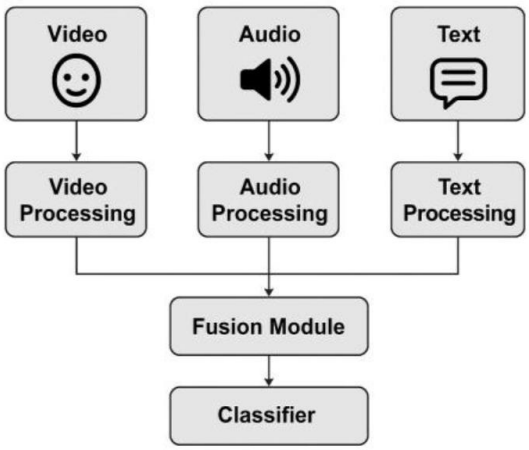

Fig. 2. Typical block diagram of the architecture of a multimodal emotion recognition system

Fig. 2 schematically illustrates a typical pipeline of a multimodal system: three input modalities pass through separate processing blocks, their features are then merged, and finally, the classifier outputs an emotion prediction.

Practical implementation considerations. Emotional classification based on multimodal data involves integrating information from different modalities–text, audio, and video [20]. The proposed approaches differ in their fusion strategies (early, late, or hybrid) and their ability to process certain types of data.

Table 1 presents a comparative overview of leading transformer-based multimodal models focused on emotion analysis or related tasks. For each model, the modalities involved, the type of feature fusion, the achieved accuracy values (F1 or Accuracy based on available data), key architectural features, and limitations are indicated.

Table 1 – Key features and performance of modern multimodal models for emotional classification tasks

Model Name	Modal	Fusion	Accuracy
Adapted Multimod BERT (AMB)	Text + Audio / Video	Hybrid (Layer-wise)	84.2 %
Flamingo	Image / Video + Text	Hybrid (attention)	78.3 %
SpeechT5	Text + Audio	Hybrid (encod)	76.5 % WER
MMBT	Text + Image	Early	92.4 %
Video BERT	Text + Video	Mixed	52.1 %

The Accuracy column lists the Accuracy results according to the best available data; Features and Limitations describe the architectural approaches and limitations of the models.

Analysis shows that hybrid architectures, which combine modality-specific processing and joint training, such as MultimodalBERT or SpeechT5, yield the most balanced results in terms of accuracy and flexibility. In contrast, high-accuracy models such as MMBT are less versatile and require separate processing of input features. This highlights the typical trade-off between efficiency, scalability, and versatility in multimodal approaches.

To visually compare the effectiveness of different multimodal architectures, an accuracy chart was constructed based on publicly available model test results on relevant tasks.

As shown in Fig. 3, the MMBT model, which employs projective fusion of visual and textual features, achieves the highest accuracy, yielding a result of 91.2 % on the meme classification task. The AMB model with a Hybrid architecture also demonstrates high performance in multimodal emotion classification, achieving an accuracy of 84.2 % on the CMU-MOSEI dataset.

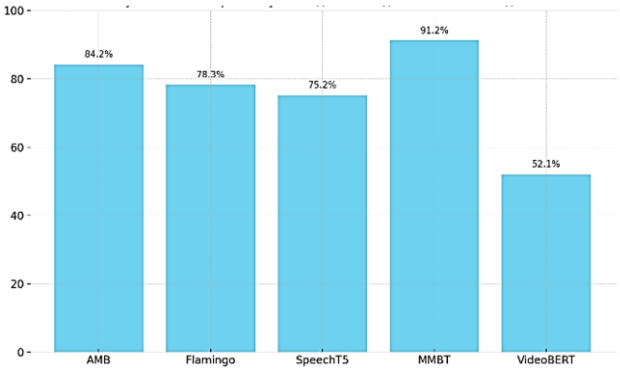

Fig. 3. Accuracy of different multimodal models on test tasks

In contrast, general-purpose architectures such as VideoBERT, Flamingo, and SpeechT5 are less accurate, partly because they are designed for general multimodal tasks rather than specialized emotional scenarios. The results obtained emphasize the importance of adapting fusion mechanisms to the nature of the input data and the target task.

Conclusions. This article provides a systematic review of methods for processing and integrating audio, video, and text data using artificial intelligence. It has been established that multimodal systems demonstrate a significant increase in the accuracy of classification and information interpretation compared to single-channel approaches, provided that the modalities are properly synchronized and relevant features are identified.

The main integration architectures (early, late, and hybrid fusion) are described, and a comparative analysis of their properties is performed. The scientific novelty of the work lies in its comprehensive systematization of architectures. It employs multimodal analysis, which considers current trends in large-scale pre-trained transformer models with cross-modal attention mechanisms. Conceptual schemes for adaptive data fusion are proposed, which highlight registers of modality-specific features and combine them, taking into account cross-modal relevance.

The practical value lies in the formulation of recommendations for designing robust multimodal systems across various application domains, taking into account their accuracy and adaptability.

The limitations of the current analysis include its focus on three primary modalities (audio, video, and text), as well as the requirement for substantial amounts of annotated data and computational resources for model training. The asynchrony and heterogeneity of input signals

complicate the direct combination of features, requiring specific preprocessing and synchronization methods. Further research should focus on developing hybrid multimodal models with dynamic adaptation of fusion schemes and cross-modal attention mechanisms, as well as on experimentally testing their effectiveness in real-world tasks.

References

1. Yuan Y., Li Z., Zhao B. A Survey of Multimodal Learning: Methods, Applications, and Future. *ACM Computing Surveys*. 2025. Vol. 57, no. 7. P. 1–34. DOI: 10.1145/3713070.
2. Golovanevsky M., Schiller E., Nair A., Han E., Singh R., Eickhoff C. One-Versus-Others Attention: Scalable Multimodal Integration for Biomedical Data. *Pacific Symposium on Biocomputing 2025: Biocomputing 2025*. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC, 2024. P. 580–593. DOI: 10.1142/9789819807024_0041.
3. Xue J., Wang Y., Tian Y., Li Y., Shi L., Wei L. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*. 2021. Vol. 58, no. 5. P. 102610-102624. DOI: 10.1016/j.ipm.2021.102610.
4. Xie Y., Yang L., Zhang M., Chen S., Li J. A Review of Multimodal Interaction in Remote Education: Technologies, Applications, and Challenges. *Applied Sciences*. 2025. Vol. 15, no. 7. P. 3937–3964. DOI: 10.3390/app15073937.
5. Wilson A., Wilkes S., Teramoto Y., Hale S. Multimodal analysis of disinformation and misinformation. *Royal Society Open Science*. 2023. Vol. 10, no. 12. P. 230964–230989. DOI: 10.1098/rsos.230964.
6. Alaba S. Y., Gurbuz A. C., Ball J. E. Emerging Trends in Autonomous Vehicle Perception: Multimodal Fusion for 3D Object Detection. *World Electric Vehicle Journal*. 2024. Vol. 15, no. 1. P. 20–30. DOI: 10.3390/wevj15010020.
7. Warner E., Lee J., Hsu W., Syeda-Mahmood T., Kahn C. E., Gevaert O., Rao A. Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects // *International Journal of Computer Vision*. 2024. Vol. 132, no. 9. P. 3753–3769. DOI: 10.1007/s11263-024-02032-8.
8. Lian H., Lu C., Li S., Zhao Y., Tang C., Zong Y. A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. *Entropy*. 2023. Vol. 25, no. 10. P. 1440–1473. DOI: 10.3390/e25101440.
9. Khan M., Tran P.-N., Pham N. T., El Saddik A., Othmani A. MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion. *Scientific Reports*. 2025. Vol. 15, no. 1. P. 5473–5486. DOI: 10.1038/s41598-025-89202-x.
10. Udahemuka G., Djouani K., Kurien A. M. Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review. *Applied Sciences*. 2024. Vol. 14, no. 17. P. 8071–8115. DOI: 10.3390/app14178071.
11. Caschera M. C., Grifoni P., Ferri F. Emotion Classification from Speech and Text in Videos Using a Multimodal Approach. *Multimodal Technologies and Interaction*. 2022. Vol. 6, no. 4. P. 28–54. DOI: 10.3390/mti6040028.
12. Tsai Y. H., Bai S., Liang P. P., Kolter J. Z., Morency L. P., Salakhutdinov R. Multimodal Transformer for Unaligned Multimodal Language Sequences. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. P. 6558–6569. DOI: 10.18653/v1/P19-1656.
13. Farhadizadeh M., Weymann M., Blaß M., Kraus J., Gundler C., Walter S., Hempen N., Binder H., Binder N. A Systematic Review of Challenges and Proposed Solutions in Modeling Multimodal Data. *arXiv*. 2025. DOI: 10.48550/ARXIV.2505.06945.
14. Wu Y., Zhang S., Li P. Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features. *Scientific Reports*. 2025. Vol. 15, no. 1. P. 8855–8888. DOI: 10.1038/s41598-025-89758-8.
15. Das A., Sarma M. S., Hoque M. M., Siddique N., Dewan M. A. A. AVaTER: Fusing Audio, Visual, and Textual Modalities Using Cross-Modal Attention for Emotion Recognition. *Sensors*. 2024. Vol. 24, no. 18. P. 5862–5886. DOI: 10.3390/s24185862.
16. Xu P., Zhu X., Clifton D. A. Multimodal Learning with Transformers: A Survey. *arXiv*. 2023. DOI: 10.48550/arXiv.2206.06488.
17. Alayrac J. B., Donahue J., Luc P., Miech A., Barr I. та ін. A Visual Language Model for Few-Shot Learning. *arXiv*. 2022. DOI: 10.48550/ARXIV.2204.14198.
18. Sun C., Myers A., Vondrick C., Murphy K., Schmid C. VideoBERT: A Joint Model for Video and Language Representation Learning. *arXiv*. 2019. DOI: 10.48550/ARXIV.1904.01766.
19. Sun Z., Lin M., Zhu Q., Xie Q., Wang F., Lu Z., Peng Y. A scoping review on multimodal deep learning in biomedical images and texts. *Journal of Biomedical Informatics*. 2023. Vol. 146. P. 104482–104502. DOI: 10.1016/j.jbi.2023.104482.
20. Kaczmarczyk R., Wilhelm T. I., Martin R., Roos J. Evaluating multimodal AI in medical diagnostics. *Digital Medicine*. 2024. Vol. 7, no. 1. P. 205–210. DOI: 10.1038/s41746-024-01208-3.

References (transliterated)

1. Yuan Y., Li Z., Zhao B. A Survey of Multimodal Learning: Methods, Applications, and Future. *ACM Computing Surveys*. 2025. Vol. 57, no. 7, pp. 1–34. DOI: 10.1145/3713070.
2. Golovanevsky M., Schiller E., Nair A., Han E., Singh R., Eickhoff C. One-Versus-Others Attention: Scalable Multimodal Integration for Biomedical Data. *Pacific Symposium on Biocomputing 2025: Biocomputing 2025*. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC, 2024, pp. 580–593. DOI: 10.1142/9789819807024_0041.
3. Xue J., Wang Y., Tian Y., Li Y., Shi L., Wei L. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*. 2021, vol. 58, no. 5, pp. 102610-102624. DOI: 10.1016/j.ipm.2021.102610.
4. Xie Y., Yang L., Zhang M., Chen S., Li J. A Review of Multimodal Interaction in Remote Education: Technologies, Applications, and Challenges. *Applied Sciences*. 2025, vol. 15, no. 7, pp. 3937–3964. DOI: 10.3390/app15073937.
5. Wilson A., Wilkes S., Teramoto Y., Hale S. Multimodal analysis of disinformation and misinformation. *Royal Society Open Science*. 2023, vol. 10, no. 12, pp. 230964–230989. DOI: 10.1098/rsos.230964.
6. Alaba S. Y., Gurbuz A. C., Ball J. E. Emerging Trends in Autonomous Vehicle Perception: Multimodal Fusion for 3D Object Detection. *World Electric Vehicle Journal*. 2024, vol. 15, no. 1, pp. 20–30. DOI: 10.3390/wevj15010020.
7. Warner E., Lee J., Hsu W., Syeda-Mahmood T., Kahn C. E., Gevaert O., Rao A. Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects. *International Journal of Computer Vision*. 2024, vol. 132, no. 9, pp. 3753–3769. DOI: 10.1007/s11263-024-02032-8.
8. Lian H., Lu C., Li S., Zhao Y., Tang C., Zong Y. A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. *Entropy*. 2023, vol. 25, no. 10, pp. 1440–1473. DOI: 10.3390/e25101440.
9. Khan M., Tran P.-N., Pham N. T., El Saddik A., Othmani A. MemoCMT: multimodal emotion recognition using cross-modal transformer-based feature fusion. *Scientific Reports*. 2025, vol. 15, no. 1, pp. 5473–5486. DOI: 10.1038/s41598-025-89202-x.
10 Udahemuka G., Djouani K., Kurien A. M. Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review. *Applied Sciences*. 2024, vol. 14, no. 17, pp. 8071–8115. DOI: 10.3390/app14178071.
11/ Caschera M. C., Grifoni P., Ferri F. Emotion Classification from Speech and Text in Videos Using a Multimodal Approach. *Multimodal Technologies and Interaction*. 2022, vol. 6, no. 4, pp. 28–54. DOI: 10.3390/mti6040028.
12. Tsai Y. H., Bai S., Liang P. P., Kolter J. Z., Morency L. P., Salakhutdinov R. Multimodal Transformer for Unaligned Multimodal Language Sequences. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 6558–6569. DOI: 10.18653/v1/P19-1656.
13. Farhadizadeh M., Weymann M., Blaß M., Kraus J., Gundler C., Walter S., Hempen N., Binder H., Binder N. A Systematic Review of Challenges and Proposed Solutions in Modeling Multimodal Data. *arXiv*. 2025. DOI: 10.48550/ARXIV.2505.06945.
14. Wu Y., Zhang S., Li P. Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features. *Scientific Reports*. 2025, vol. 15, no. 1, pp. 8855–8888. DOI: 10.1038/s41598-025-89758-8.

15. Das A., Sarma M. S., Hoque M. M., Siddique N., Dewan M. A. A. AVaTER: Fusing Audio, Visual, and Textual Modalities Using Cross-Modal Attention for Emotion Recognition. *Sensors*. 2024, vol. 24, no. 18, pp. 5862–5886. DOI: 10.3390/s24185862.
16. Xu P., Zhu X., Clifton D. A. Multimodal Learning with Transformers: A Survey. *arXiv*. 2023. DOI: 10.48550/arXiv.2206.06488.
17. Alayrac J. B., Donahue J., Luc P., Miech A., Barr I. та ін. A Visual Language Model for Few-Shot Learning. *arXiv*. 2022. DOI: 10.48550/ARXIV.2204.14198.
18. Sun C., Myers A., Vondrick C., Murphy K., Schmid C. VideoBERT: A Joint Model for Video and Language Representation Learning. *arXiv*. 2019. DOI: 10.48550/ARXIV.1904.01766.
19. Sun Z., Lin M., Zhu Q., Xie Q., Wang F., Lu Z., Peng Y. A scoping review on multimodal deep learning in biomedical images and texts. *Journal of Biomedical Informatics*. 2023, vol. 146, pp. 104482–104502. DOI: 10.1016/j.jbi.2023.104482.
20. Kaczmarczyk R., Wilhelm T. I., Martin R., Roos J. Evaluating multimodal AI in medical diagnostics. *Digital Medicine*. 2024, vol. 7, no. 1, pp. 205–210. DOI: 10.1038/s41746-024-01208-3.

Received 15.11.2025

УДК 004.62

О. В. ЖЕРЕБЕЦЬКИЙ, аспірант кафедри Систем штучного інтелекту, Національного університету «Львівська політехніка», м. Львів, Україна; e-mail: oleh.v.zherebetskyi@lpnu.ua; ORCID: https://orcid.org / 0009-0004-6259-7065
О. А. БАСИСТЮК, кандидат технічних наук (PhD), старший викладач кафедри Систем штучного інтелекту, Національного університету «Львівська політехніка», м. Львів, Україна; e-mail: oleh.a.basystiuk@lpnu.ua; ORCID: https://orcid.org /0000-0003-0064-6584

КОМПЛЕКСУВАННЯ РІЗНОТИПОВИХ ДАНИХ ЗАСОБАМИ ШТУЧНОГО ІНТЕЛЕКТУ

У сучасній розробці штучного інтелекту методи мультимодального аналізу даних набувають критичного значення завдяки своїй здатності інтегрувати інформацію з різних джерел, включаючи текст, аудіо, сигнали датчиків та зображення. Така інтеграція дозволяє системам формувати багатше та контекстно-залежне розуміння складних середовищ, що є важливим для таких галузей, як діагностика охорони здоров'я, адаптивні освітні технології, інтелектуальні системи безпеки, автономна робототехніка та різні форми взаємодії людини з комп'ютером. Мультимодальні підходи також дозволяють моделям ШІ компенсувати обмеження, властиві окремим модальностям, тим самим підвищуючи стійкість та стійкість до шуму або неповних даних. У дослідженні використовується теоретичний аналіз наукової літератури, порівняльна класифікація мультимодальних архітектур, систематизація методів об'єднання та формальне узагальнення принципів проектування моделей. Крім того, увага приділяється оцінці нових парадигм, що базуються на великомасштабних фундаментальних моделях та архітектурах на основі трансформаторів. Узагальнено основні методи та моделі обробки мультимодальних даних, що охоплюють як класичні, так і найсучасніші підходи. Архітектури раннього (на рівні ознак), пізнього (на рівні рішень) та гібридного (проміжного) об'єднання описані та порівняні з точки зору гнучкості, обчислювальної складності, інтерпретованості та точності. Також аналізуються нові рішення, засновані на великих мультимодальних трансформаторних моделях, контрастному навчанні та уніфікованих просторах вбудовування. Особлива увага приділяється механізмам крос-модальної уваги, які дозволяють динамічне зважування модальностей залежно від контексту завдання. Дослідження визначає, що мультимодальні системи досягають значно вищої точності, стабільності та семантичної узгодженості в завданнях класифікації, виявлення та інтерпретації, коли модальності належним чином синхронізовані та об'єднані за допомогою адаптивних стратегій. Ці результати підкреслюють перспективність подальших досліджень у напрямку масштабованих архітектур, здатних до мультимодального мислення в реальному часі, покращеного крос-модального перенесення та контекстно-залежних механізмів уваги.

Ключові слова: мультимодальність, штучний інтелект, емоційна класифікація, ф'южн-архітектури, обробка аудіо-відео-тексту, трансформери, крос-модальна увага.

Повні імена авторів / Author's full names

Автор 1 / Author 1: Жеребецький Олег В'ячеславович / Zherebetskyi Oleh Vyacheslavovych
Автор 2 / Author 2: Басистюк Олег Андрійович / Basystiuk Oleh Andriyovych