https://doi.org/10.31891/csit-2026-1-3

UDC 004.9

DEVELOPMENT AND RESEARCH OF MULTIMODAL NEURAL ARCHITECTURES FOR -HETEROGENEOUS UNBALANCED DATA IN CLASSIFICATION TASKS

Serhii MINUKHIN
Professor of the Department of Computer Modeling and Intelligent Technologies?
Professor of the Department of Information Systems,
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
serhii.minukhin@nure.ua
Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine
serhii.minukhin@hneu.net
https://orcid.org/0000-0002-9314-3750

Valerii RUDOI
PhD Student
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
valerii.rudoi@nure.ua
https://orcid.org/0009-0002-5285-7746

The article presents a comprehensive study of modern multimodal neural architectures for integrating heterogeneous and partially unbalanced data in classification tasks. It considers early and late fusion approaches, hybrid architectures with cross-modal attention, and transformers that allow the formation of consistent latent spaces of visual, auditory, and textual features. Particular attention is paid to contrastive learning (CLIP-like approaches, multimodal InfoNCE), which ensures semantic consistency of representations and improves classification accuracy in the presence of uneven data distribution and rare classes. A model is proposed that combines early and late fusion with cross-modal attention and contrastive learning to form a coherent joint latent space. Features of each modality are processed by specialized encoders, and fusion is performed with adaptive weighting, which minimizes the impact of heterogeneous data imbalance and enables the efficient processing of signals of different natures and intensities. The use of pruning, quantization, and knowledge distillation has reduced computational costs without losing accuracy, ensuring stable model performance in real-world streaming scenarios with limited resources. The results of applying the proposed model to the BDD100K and CMU-MOSEI datasets confirmed the model's high efficiency in processing heterogeneous and unbalanced data. For BDD100K, Accuracy 0.953, F1-score 0.956, ROC-AUC 0.947 were achieved, and the integral indicators Micro F1, Macro F1, and Weighted F1 were 0.953, 0.949, and 0.955, respectively; For CMU-MOSEI, Accuracy 0.956, F1-score 0.969, ROC-AUC 0.968, and the integral indicators Micro F1, Macro F1, and Weighted F1 were 0.956, 0.962, and 0.968, respectively. A comparative analysis with classical feature concatenation approaches, recent State-of-the-Art multimodal fusion models, and AutoML-based solutions demonstrated that the proposed architecture consistently outperforms existing methods. In particular, the model improves classification accuracy by approximately 2–4% compared to recent SOTA architectures and provides more stable F1-scores for minority classes. A comparison with the AutoML-based framework B-T4SA also confirms the robustness of the proposed approach. These results demonstrate that the developed model ensures higher classification consistency for both frequent and rare classes under heterogeneous and imbalanced data conditions.

Keywords: multimodal data, cross-modal attention, contrastive learning, knowledge distillation, pruning, quantization, emotion classification, autonomous navigation.

Received: 07/01/2026
Accepted:14/02/2026
Published:26/03/2026

Introduction

Multimodal data analysis has emerged as a pivotal research area within contemporary artificial intelligence, as it facilitates the integration of heterogeneous information sources, encompassing visual, auditory, textual, video, and sensor-based signals. The joint processing of multiple modalities enables AI systems to construct enriched, semantically coherent, and context-aware representations of complex real-world phenomena. This capability is particularly critical in dynamic and high-uncertainty domains such as autonomous driving, medical diagnostics, human-computer interaction, and emotion recognition, where reliance on a single modality frequently yields incomplete or unreliable outcomes, potentially compromising decision-making processes or system safety [1].

Despite its significant potential, multimodal learning presents substantial methodological challenges arising from intrinsic inter-modal heterogeneity. Individual modalities may exhibit distinct statistical distributions, variable temporal resolutions, differing noise characteristics, and unique semantic structures. Such disparities often result in feature misalignment, where relevant patterns within one modality are not appropriately correlated with complementary information from other modalities. Furthermore, modality dominance may occur, wherein the learning process is biased toward stronger modalities at the expense of weaker but informative ones. Conflicting information across modalities may also introduce ambiguity, thereby degrading overall model performance. These factors complicate the design of fusion strategies that are simultaneously robust and capable of capturing complex cross-modal relationships.

Conventional fusion approaches, including early fusion, late fusion, and simple feature concatenation, frequently prove insufficient in addressing these challenges. Early fusion methods, which integrate raw features from multiple modalities, are sensitive to scale discrepancies and noise, potentially distorting the joint representation. Late fusion strategies, which combine predictions from modality-specific models, often neglect fine-grained inter-modal interactions [2]. Likewise, feature concatenation does not guarantee semantic consistency, limiting the model's ability to fully exploit complementary information across modalities. Consequently, such traditional approaches often underperform on tasks that require a nuanced understanding of cross-modal dependencies.

Recent advancements in deep learning provide promising avenues to mitigate these limitations. Architectures leveraging attention mechanisms, shared latent representations, contrastive learning frameworks, and transformer-based models have demonstrated improved feature alignment, enhanced robustness to noise, and resilience to missing data. However, these approaches typically require large-scale annotated datasets and significant computational resources, along with careful hyperparameter tuning and architectural optimization. This poses challenges in resource-constrained settings or in applications that demand real-time processing. Accordingly, a central research challenge in multimodal AI lies in developing strategies that optimally balance representational accuracy, computational efficiency, and scalability, while maintaining semantically consistent and robust cross-modal representations [3].

The remainder of this paper is organized as follows. The section Related Works reviews existing research on multimodal learning architectures and fusion strategies. The section Purpose and Objectives of the Study formulates the main aim of the research and the specific tasks addressed in this work. The section Materials and Methods describes the theoretical foundations of multimodal integration, the proposed hybrid architecture, and optimization techniques. The sections Datasets and Data Preprocessing present the characteristics of the selected datasets and the procedures applied to prepare the multimodal data for training. The section Experiments and Results reports the experimental evaluation and comparative analysis of the proposed model. Finally, the section Conclusions, Limitations, and Future Work summarizes the key findings and outlines directions for further research.

Related works

Multimodal learning is a well-established research field that focuses on integrating heterogeneous data sources, including images, text, audio, sensor signals, and structured features, to improve the quality of representations and predictive performance [1–4]. Foundational research identifies three main fusion strategies: early, intermediate, and late fusion [5–8]. Early fusion allows direct modeling of cross-modal dependencies at the feature level and creates a unified representation of all modalities, but it is sensitive to differences in scale, noise, and statistical properties between sources [6,9]. Intermediate fusion, based on shared or partially shared latent spaces, has become the dominant paradigm in modern neural architectures, as it can capture complex interdependencies while preserving modality-specific characteristics [7,10–12]. Late fusion aggregates outputs or features from independently trained subnetworks, which is particularly effective for incomplete, asynchronous, or highly heterogeneous data and supports model interpretability [13–15].

To summarize the main characteristics of existing approaches, Table 1 presents a comparison of multimodal fusion strategies, modalities used in typical applications, and their primary limitations.

Table 1

Comparison of multimodal fusion strategies

Fusion type	Typical modalities	Main idea	Advantages	Limitations
Early fusion	Image + text, image + sensor data	Concatenation or joint encoding of features before model training	Captures direct cross-modal interactions	Sensitive to noise, scale differences, and missing modalities
Intermediate fusion	Image + text + audio, multimodal embeddings	Learning shared latent representations using neural networks	Preserves modality-specific features while modeling interactions	Requires complex architectures and careful training
Late fusion	Independent modality-specific models	Combining predictions or high-level features from separate models	Robust to missing or asynchronous modalities, interpretable	Limited ability to learn deep cross-modal relationships

Comprehensive surveys of multimodal deep learning highlight architectures based on convolutional, recurrent, and transformer networks, emphasizing attention mechanisms, multi-task learning, and pretraining strategies for improved feature alignment and generalization [16–19]. Interpretable heterogeneous ensembles represent a complementary direction, balancing predictive performance and transparency, which is particularly important in biomedical and healthcare applications [14,17,18].

In practical applications, multimodal networks demonstrate significant improvements in healthcare and finance. In healthcare, they integrate medical imaging, genomic profiles, and clinical records to enhance diagnostic accuracy, predict disease progression, and support personalized treatment planning [1,3,4]. In finance, combining numerical indicators, textual news, and social media sentiment enables better risk assessment, trend forecasting, and investment optimization [2,4].

Two of the most intensively studied domains are multimodal sentiment analysis and autonomous systems. In sentiment analysis, AutoML-based fusion strategies achieve state-of-the-art performance, with the B-T4SA model reaching an accuracy of 95.19% [19]. In autonomous driving, multimodal models integrate camera images, LiDAR, radar, sensor data, and semantic scene descriptions. Large-scale benchmarks, such as BDD100K, facilitate heterogeneous multi-task learning and support explainable frameworks for perception and decision-making in complex environments [3,4,19].

Despite these advances, several important limitations remain in existing multimodal learning approaches. First, traditional fusion strategies such as early and late fusion often fail to ensure deep semantic alignment between heterogeneous modalities, which may lead to information loss or insufficient exploitation of complementary features. Second, many modern architectures rely on large-scale labeled datasets and high computational resources, making them difficult to deploy in real-time or resource-constrained environments. Third, existing models frequently struggle with modality imbalance, where dominant modalities overshadow weaker but informative signals, reducing the robustness of multimodal representations. Finally, many current solutions provide limited flexibility in adapting fusion mechanisms to heterogeneous data conditions or dynamically adjusting the contribution of individual modalities. These limitations highlight the need for more flexible and computationally efficient multimodal architectures capable of achieving stable feature alignment, balanced modality integration, and high predictive performance across diverse application domains.

Purpose and objectives of the study

The aim of this research is to conduct a comparative analysis of modern architectures for multimodal learning, specifically classical and hybrid systems, and to evaluate their effectiveness in terms of accuracy, energy efficiency, and scalability metrics.

To achieve this purpose, the following objectives must be solved:

1. Conduct a theoretical analysis of modern approaches to multimodal integration, including early and late fusion mechanisms, as well as methods for matching features in a shared latent space;
2. Justify the choice of metrics for evaluating the effectiveness of multimodal models, taking into account their architectural features, computational complexity, and resistance to input data variability;
3. Prepare and perform pre-processing of two multimodal datasets representing visual-sensory scenes of the road environment and audiovisual-textual emotional interactions, respectively, ensuring the unification of data formats and bringing all modalities to a form suitable for joint processing in neural networks;
4. Conduct experimental research and comparative analysis of different multimodal fusion strategies based on convolutional, recurrent, and transformer architectures, analyze the results, and form conclusions about their effectiveness.

Materials and methods

The traditional paradigm of multimodal integration has long been based on two basic approaches: early and late fusion. Early fusion involves combining features from different modalities before they are passed into the main neural network. This approach allows the creation of a single latent representation that already contains information from all available sources. Formally, this process is described by the equation (see formula 1) [4]:

$$Z = f_{\text{fusion}}(x_1, x_2, \ldots, x_M), \tag{1}$$

where x_i – input data if the i-th modality, Z – integrated representation, M – number of modalities. The advantage of early integration is deeper feature alignment, since all modalities are presented in a common space before processing begins, allowing learning algorithms to identify interdependencies between heterogeneous data and create more complex, multidimensional representations.

Late fusion, on the contrary, involves independent training of subnetworks $\phi_i(x_i, W_i)$with subsequent aggregation through the function (see formula 2) [5]:

$$\hat{y} = g(\phi_1(x_1, W_1), \ldots, \phi_M(x_M, W_M)). \tag{2}$$

This approach allows for preserving the specificity of each modality, supporting the independent training and adaptation of subnetworks to changes in the statistical properties of the corresponding modal data or conditions of their arrival. Additionally, it greatly simplifies the addition of new data sources without the need to retrain the entire system. It reduces the risk of “noise mixing” from heterogeneous sources, although it may limit the depth of semantic alignment.

With the advent of transformers, it has become possible to flexibly represent dependencies between all elements of different modalities through a self-attention mechanism.

The classical formalization of attention is represented as follows (see formula 3) [6]:

$$\text{Attention}(Q_i, K_j, V_j) = \text{softmax}\left(\frac{Q_i K_j}{\sqrt{d_k}}\right) V_j, \tag{3}$$

where Q_i, K_j, V_j – matrices of queries, keys, and values of the corresponding modalities, d_k – dimension of the keys. This enables the implementation of cross-modal connections that encompass not only text and images, but also any combination of data, such as video with sensory signals or audio with time series [7].

For multi-headed cross-modal attention, the output representations of each layer can be aggregated with adaptive weight α_l (see formula 4) [8]:

$$Z_{\text{multi}} = \sum_{l=1}^{L} \alpha_l \cdot \text{Attention}(Q_i, K_j, V_j). \tag{4}$$

The contrastive approach to reconciling heterogeneous data is formalised through the loss function (see formula 5) [9]:

$$L_{\text{constrastive}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(z_i^{\text{text}}, z_i^{\text{image}})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(z_i^{\text{text}}, z_i^{\text{image}})/\tau)}, \tag{5}$$

where $\text{sim}()$ – cosine similarity, τ – temperature coefficient, N– package batch size.

This approach enables the model to form semantically consistent representation spaces, where objects or events with similar content or belonging to the same class are positioned closer together in the vector space, while incompatible or heterogeneous objects are moved further apart. This representation structure enhances the model's performance in various tasks, including classification, information retrieval, example comparison, and the generation of new content based on integrated features.

Contrastive loss functions are particularly effective when working with large datasets and multimodal scenarios, where data comes from different sources and contains heterogeneous characteristics. They ensure global alignment of semantic information across all examples in a batch, allowing the model to identify common features and significant patterns regardless of modality.

Network architecture optimization is performed through Neural Architecture Search (NAS), where the loss function on validation data is minimized depending on the architecture parameters (see formula 6) [10]:

$$\min_{\alpha, W} L_{\text{val}}(W^*(\alpha), \alpha), W^*(\alpha) = \arg\min_{W} L_{\text{train}}(W, \alpha). \tag{6}$$

It involves a two-stage optimization process: first, the optimal network weights for the given architecture parameters are determined, and then the architecture parameters that minimize the loss function on the validation data are selected. Thus, NAS allows you to automatically determine feature fusion points, the number and configuration of layers, and the optimal weight ratios between modalities.

In the present study, NAS was applied to determine the optimal configuration of the multimodal neural architecture. The search space included several architectural parameters that significantly affect model performance and computational complexity. In particular, the following parameters were optimized: the number of layers in modality-specific subnetworks (from 2 to 6 layers), the dimensionality of hidden feature representations (128, 256, and 512 units), and the type of activation functions (ReLU, GELU, and LeakyReLU). In addition, the search process considered the number of attention heads in cross-modal attention modules (from 4 to 8), the dropout rate (0.1–0.4), and the positions of fusion layers responsible for integrating information from different modalities.

The architecture search was performed using an iterative optimization strategy based on validation loss minimization. During the search process, 120 candidate architectures were generated and evaluated. Each candidate configuration was trained for a limited number of epochs on the training dataset and subsequently evaluated on the validation set using the cross-entropy loss and the F1-score as evaluation criteria. The architecture demonstrating the lowest validation loss and stable generalization performance was selected as the final configuration used in subsequent experiments.

Computational optimization of neural networks encompasses key methods, including pruning, quantization, and knowledge distillation, which strike a balance between performance, accuracy, and computational costs (Table 2).

Table 2

Methods for improving the efficiency of multimodal models

Method	Approach Description	Impact on Accuracy	Impact on Computational Cost	Main Advantages	Main Limitations
Pruning	Removal of insignificant weights, neurons, or connections	Accuracy may slightly decrease under aggressive pruning	Significant reduction in the number of parameters and FLOPs	Model size reduction, faster inference	Requires careful tuning and retraining
Quantization	Reducing the bit-width of weights and activations (FP32 → INT8, etc.)	Slight decrease, usually controllable	Reduced memory and energy consumption	High efficiency on edge devices	Sensitive to data type and hardware support
Knowledge Distillation	Transferring knowledge from a large model (teacher) to a compact model (student)	Often maintained or even improved	Significantly lower inference cost	High accuracy-to-size ratio	Additional cost for training the teacher model

Pruning is the removal of low-impact weights to reduce model size and computational complexity during training and inference. This approach identifies and eliminates parameters whose contribution to the output is minimal, resulting in a more compact and efficient network while preserving predictive accuracy. In large-scale multimodal systems, pruning significantly reduces memory usage and computational cost. When combined with quantization and knowledge distillation, it enables the development of lightweight models that maintain semantic consistency and high performance under limited computational resources.

Quantization converts weights into discrete levels, which reduces memory consumption and speeds up computation, while knowledge distillation allows knowledge to be transferred from a large "teacher" model to a compact 'student' model by minimizing Kullback-Leibler divergence (see formula 7) [11]:

$$L_{\text{KD}} = \sum_i p_i^{\text{teacher}} \log \frac{p_i^{\text{teacher}}}{p_i^{\text{student}}}, \tag{7}$$

where p_i^{teacher} represents the probability of class i, predicted by the large model, while p_i^{student} represents analogous prediction of the smaller model.

In the proposed framework, the teacher model corresponds to the full multimodal architecture obtained after the NAS optimization stage. This model includes modality-specific subnetworks combined through transformer-based cross-modal attention mechanisms and operates with high-dimensional feature representations (512 units). Owing to its structural complexity and large number of parameters, the teacher model achieves high predictive accuracy but requires substantial computational resources.

The student model represents a compressed version of the teacher architecture designed for computational efficiency. It contains a reduced number of layers in each modality branch (3–4 layers) and employs lower-dimensional latent representations (256 units). During training, the student network is optimized using a combination of the standard classification loss and the distillation loss derived from the soft probability distributions produced by the teacher model.

Knowledge distillation losses quantify how accurately the student model reproduces the behavior of the teacher model, enabling effective knowledge transfer from a large network to a compact one with minimal performance degradation. By learning from both hard labels and the teacher's soft predictions, the student preserves the semantic structure and inter-class relationships encoded in the teacher's representations. This approach is particularly effective for large-scale multimodal models, allowing the construction of computationally efficient student networks that maintain semantic consistency across text, visual, audio, and video modalities while capturing complex hidden features and contextual dependencies.

A balance between modalities is necessary to compensate for the dominance of one modality in the final representation. The loss function in this case is as follows (see formula 8) [12]:

$$L_{\text{multi}} = \sum_{i=1}^{M} \lambda_i L_i \,, \sum_{i=1}^{M} \lambda_i = 1, \tag{8}$$

where L_i – losses calculated for each modality separately, λ_i – integration weight coefficients.

This approach enables the dynamic adjustment of the influence of each modality during training or inference, which is particularly important in cases where one modality dominates in terms of information value, thereby ensuring a balance between data sources and preventing any one modality from dominating over others.

Summarizing the various approaches to optimizing multimodal models, it is advisable to use an integrated loss function that combines classification accuracy and the alignment of features from different modalities in a shared space. Such a loss function is given by the formula (see formula 9) [13]:

$$L_{\text{embedding}} = L_{CE} + \lambda \cdot L_{\text{align}}, \tag{9}$$

where L_{CE} – cross-entropy loss, L_{align} – alignment loss, which estimates how well the features of different modalities align in the shared embedding space, λ – weighting coefficient that balances the contribution of alignment and classification accuracy to the overall loss function.

Small values λ lead to the dominance of cross-entropy loss and, accordingly, to weak intermodal consistency, while excessively large values reduce the discriminative power of the classifier by reorienting the optimization towards geometric alignment of features.

The optimal value λ is defined as the one that ensures stable growth of generalization metrics on the validation set without degrading classification accuracy. In the experiments conducted, the selection λ was made taking into account the criterion of balance between the F1-score and the amount of alignment loss, which allowed us to achieve a consistent representation of modalities while maintaining high classification quality.

The selection of λ value $\lambda = 0.3$ is justified by empirical analysis on the validation set, which showed that this value provides the optimal balance between cross-entropy loss and intermodal alignment loss. At lower values λ insufficient alignment of features of different modalities in the common embedding space was observed, while increasing the value above 0.3 led to a gradual decrease in the F1-score due to the loss of the classifier's discriminatory power.

This research proposes a hybrid model of multimodal integration that **combines early and late fusion approaches** with contrastive optimization and automated NAS architecture search. The model enables the formation of a common latent space of representations for heterogeneous modalities, while preserving the specificity of each subnetwork, thereby supporting the simultaneous adaptive balancing of the contributions of individual modalities. The use of a contrastive loss function allows the model to form semantically consistent feature vectors, where similar objects are brought closer together, and heterogeneous ones are moved further apart, increasing the accuracy of integration. Computational efficiency is achieved by combining pruning, quantization, and knowledge distillation, which reduces memory and computation without significant performance loss. Thanks to this comprehensive optimization, the model provides high-precision multimodal integration, effective feature alignment, and scalability on large and heterogeneous datasets.

Datasets.

To evaluate the performance of multimodal models, two datasets were selected that represent different aspects of complex signal processing, allowing for the assessment of the effectiveness of integrating heterogeneous information sources. The BDD100K and CMU-MOSEI datasets were selected for experimental testing of multimodal architectures due to their representativeness and heterogeneity of modalities. BDD100K is characterized by a large amount of visual data for autonomous navigation tasks in various realistic conditions, which allows evaluating the stability of models to variability and noise. CMU-MOSEI encompasses audio, text, and visual modalities for emotion classification, rendering it suitable for evaluating models' ability to integrate heterogeneous information and form coherent representations. Together, these datasets cover key aspects of multimodal learning and allow for comparing the effectiveness of architectures in terms of accuracy, energy efficiency, and scalability flexibility.

The BDD100K dataset [14] is one of the largest publicly available datasets for research in the field of autonomous driving, comprising 100,000 video clips of road scenes with a total duration exceeding 1,600 hours. The structure of the set is supplemented by a large number of annotations, containing over 2,000,000 labelled objects, including cars, pedestrians, cyclists, road signs, traffic lights, and other elements of urban infrastructure. In addition, the dataset contains more than 100,000 labelled lanes and a similar number of road scenario records describing various types of behavioral situations, such as crossing intersections, changing lanes, braking, or stopping. The video clips encompass a diverse range of outdoor conditions, featuring tens of thousands of daytime and nighttime scenes, as well as thousands of clips captured in rain or fog. This provides realistic variability in the road environment, allowing the model's robustness to changes in lighting and weather conditions to be assessed. Thanks to this structure, BDD100K is optimal for this study, as it enables us to evaluate the ability of a multimodal model to recognize different types of objects, classify road situations, and analyze behavioral patterns in the dynamic conditions of autonomous navigation.

The CMU-MOSEI dataset [15] comprises 23,453 video clips from interviews, totaling over 65 hours in duration, each accompanied by annotations of six emotional states: joy, sadness, anger, fear, disgust, and surprise. Each sample in this dataset is a complete multimodal structure, as it contains synchronized video data, audio recordings and text transcripts. The video modality contains approximately 2,300,000 frames, based on which key facial points have been identified for facial expression analysis. The audio modality covers over 65 hours of recordings containing information about intonation, timbre, pauses, and other prosodic characteristics of speech. The text transmission comprises over 300,000 sentences, which collectively contain approximately 2.5 million tokens, reflecting both the content of the statements and the linguistic characteristics of the speakers.

Data Preprocessing

Pre-processing was performed on the specified datasets.

The following standardization steps were applied to the video data:

1. All frames were converted to a uniform resolution of 224×224 pixels, which is compatible with the 3D-CNN architecture and pre-trained visual encoders, thus avoiding distortions during convolutional folding.

2. The length of video clips for each fragment was standardized to a fixed number of frames: for BDD100K, 30 frames per fragment, which corresponds to approximately 1 second of video at 30 FPS, and for CMU-MOSEI, 40 frames per sample, in order to preserve sufficient context for the analysis of facial expressions and emotional manifestations. This standardization ensures the correct batching of data and optimal performance of recurrent and transformer modules. Pixel values are normalized to the range [0,1], which avoids shifts due to different lighting conditions, weather conditions or color intensity in road scenes.

For text data, tokenization was applied using byte-pair encoding (BPE). Each utterance was broken down into subworlds, which allows rare or unknown words to be processed efficiently and reduces the size of the vocabulary without losing semantic information. For the CMU-MOSEI dataset, token timestamps were also preserved for synchronization with audio and video data, enabling transformer encoders to account for the temporal dependence between modalities. As a result, each text fragment was represented as a tensor array of indices, ready for input into the model.

The audio data was converted to a Mel-spectrogram [16], allowing the model to efficiently process the frequency and temporal characteristics of the sound.

For the BDD100K dataset, audio signals are virtually absent. To enable the model to learn correlations between motion, environmental noise, and audio-like features, synthetic motion and noise spectrograms were generated. The procedure involved three steps: extraction of vehicle motion parameters (speed, acceleration, and turn rate) and environmental signals from the dataset metadata or simulation logs,transformation of these parameters into time-varying sinusoidal waveforms modulated with stochastic Gaussian noise to emulate sensor and environmental variability, and conversion of these waveforms into Mel-spectrograms using the same 16 kHz sampling rate and 128 filter banks as used for real audio in CMU-MOSEI. This process produces spectrograms that reflect the temporal dynamics of motion and ambient noise, allowing the network to learn cross-modal patterns without relying on recorded audio.

For the CMU-MOSEI dataset, the audio was converted to a sampling rate of 16 kHz, divided into 25-ms windows with a 10-ms step, and then converted to a Mel-spectrogram with 128 filters. This preserves intonation and prosody information, which is critical for emotion classification. For further training, the spectrograms were normalized and converted to a tensor format (frequency × time × channels) compatible with 2D and 3D convolutional layers.

At the final pre-processing stage, all modalities, including video, audio, and text, were converted to a single tensor format optimized for transformer and cross-modal encoders. For each sample, the time intervals between video, audio, and text were properly aligned, which allowed the model to accurately combine information from the different channels.

The diagram of the proposed multimodal model with a temporal memory and attention mechanism is shown in Fig. 1.

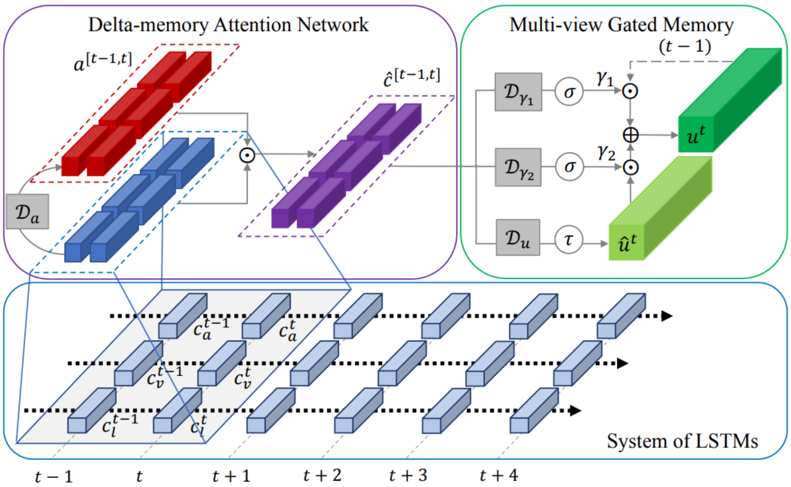

Fig. 1. Diagram of the proposed multimodal model with a temporal memory and attention mechanism

The architecture of the model is built upon a system of LSTM networks, enhanced with specialized mechanisms for processing multimodal time series. At its core lies the Delta-memory Attention Network, which captures the previous memory state between time steps $t-1$ and t, denoted as $a._{[t-1,t]}$. This attention mechanism evaluates the significance of temporal changes in the input data, selectively focusing on the most critical signals for the current processing step. Such a design enables the model to efficiently leverage historical information while emphasizing the most relevant variations in the time series.

Formally, the memory update mechanism of the Delta-memory Attention Network can be represented as a sequence of operations. Let x_t denote the multimodal input vector at time step t, and M_{t-1} denote the memory state obtained at the previous time step. First, the temporal difference between the current input and the historical memory is estimated as (see formula 10):

$$\Delta_t = x_t - M_{t-1}. \qquad (10)$$

This difference vector reflects the magnitude of change in the observed signals. Based on this variation, an attention coefficient is computed to determine the importance of the update (see formula 11):

$$\alpha_t = \sigma(W_\alpha \Delta_t + b_\alpha), \qquad (11)$$

where W_α and b_α are trainable parameters and $\sigma(\cdot)$ is the sigmoid activation function. The attention coefficient α_t controls the degree to which the current information should influence the memory update.

The new memory state is then computed as a weighted combination of the previous memory and the candidate memory representation produced by the LSTM unit (see formula 12):

$$M_t = \alpha_t \odot \tilde{M}_t - (1-\alpha_t) \odot \tilde{M}_{t-1}, \qquad (12)$$

where $\tilde{M}_t$ denotes the candidate memory produced by the LSTM cell and $\odot$ represents element-wise multiplication.

In addition, the system incorporates a Multi-view Gated Memory, which processes multiple “views” or modalities of data, denoted as D_{y1} and D_{y2}, in parallel. Each view maintains its own dedicated memory, regulated by a gating mechanism that controls the flow of information within individual channels. This arrangement allows the model to integrate heterogeneous data sources, such as video, audio, or sensor signals, while also incorporating potential external control inputs u_t that may influence the current state processing. The parameter τ serves to define a temporal interval or delay considered during the information transfer between time steps.

The network is composed of a series of LSTM blocks, represented by C_b, C_c, C_f and C_i, corresponding to different LSTM components, including input, forget, cell, and output gates. These blocks are arranged sequentially from $t-1$ to $t+4$ allowing the model to process temporal windows and account for dependencies across time. At each step, the LSTM receives multidimensional inputs Y_1, Y_2 and u_t, while the attention mechanism determines which parts of the previous memory should be activated for the current step.

The complete processing procedure for a single time step can therefore be summarized as follows:

1. Receive multimodal inputs $x_t^{(k)}$ and previous memory states $M_{t-1}^{(k)}$;
2. Compute temporal difference Δ_t and attention weights α_t;
3. Process each modality through its gated memory channel.
4. Update the LSTM internal states and generate candidate memory representations;
5. Combine the candidate and historical memories using the attention-controlled update rule.

The gated multi-view memory consolidates information across different data modalities, enabling the formation of a coherent multimodal representation. Memory states are propagated through the LSTM across time steps, ensuring the preservation of long-term dependencies. The system outputs an activated memory representation D_a along with modality-specific and control representations, denoted as D_{y1}, D_{y2} and D_u, which can subsequently be employed for classification or prediction tasks. In this way, the architecture effectively processes multimodal time series while capturing both short-term and long-term temporal dynamics.

Experiments and Results

Computational experiments were conducted in accordance with the following stages.

Stage 1. The distribution of road object classes, shooting conditions, and time of day was checked in BDD100K, and emotional classes were checked in CMU-MOSEI (Fig. 2). The results showed relative uniformity of the main categories, but in BDD100K, some object classes, such as cyclists and road signs, occur much less frequently, and some frames are blurred or contain partially obscured objects, which reduces the quality of the training data. In CMU-MOSEI, there is an imbalance between emotional classes, especially for rare categories, and the synchronization between video, audio, and text can be disrupted due to varying segment lengths.

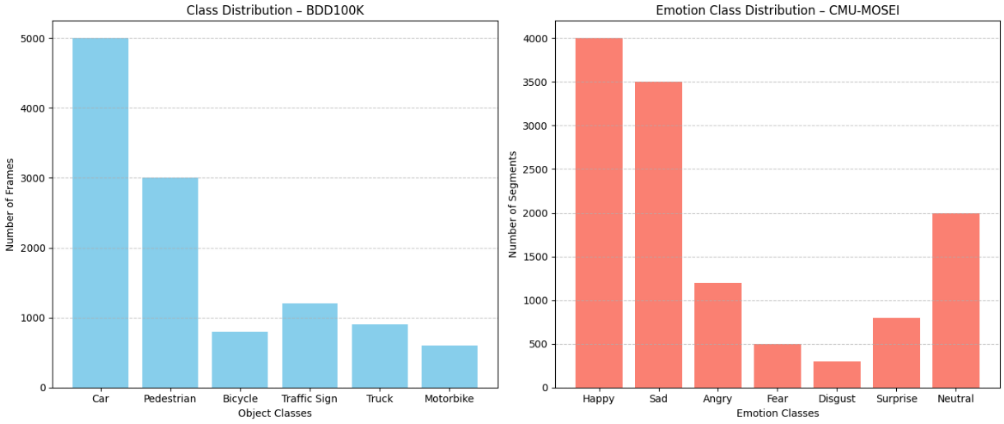

Fig. 2. Class imbalance for the BDD100K and CMU-MOSEI datasets

To solve these problems, a weighted cross-entropy loss function is used, which compensates for the underrepresentation of some classes (see formula 13) [17]:

$$L_{\text{weighted}} = -\sum_{i=1}^{C} w_i y_i \log(\bar{y}_i), \quad w_i = \frac{N_{\text{total}}}{C \cdot N_i}, \tag{13}$$

where C – total number of classes, y_i – true label of class i, $\bar{y}_i$ – predicted probability of class i, N_i – number of examples of class i, N_{total} – total number of examples in the dataset, w_i – weight coefficient of class i.

Stage 2. A unified multimodal model architecture was applied to both datasets. Video sequences were processed using 3D convolutional blocks to extract spatio-temporal features, and a recurrent module (LSTM) took into account the context of events or emotional states. For CMU-MOSEI, text and audio modalities were integrated through transformer encoders, enabling the effective combination of information from synchronized modalities.

Training was performed over 20 epochs with an adaptive learning rate schedule, and weighted cross-entropy was selected as the loss function to compensate for class imbalance. Data standardization and normalization were applied to all modalities, ensuring stable model convergence and preventing overfitting.

The key accuracy metric and F1-score for both datasets after training completion are shown in Table 3.

To assess the statistical significance of observed improvements, all performance metrics reported in Tables 2 and 3 were supplemented with 95% confidence intervals calculated using bootstrap resampling (1000 repetitions) and validated using paired t-tests. For instance, the proposed model's overall Accuracy on BDD100K (0.953) has a 95% CI of [0.950, 0.956], while CMU-MOSEI Accuracy (0.956) has a 95% CI of [0.952, 0.960], confirming that the improvements over classical methods are statistically significant ($p < 0.01$). Similar confidence intervals were obtained for F1, Precision, Recall, and ROC metrics.

Table 3

Class-oriented evaluation of the effectiveness of a multimodal model for autonomous navigation and emotional classification tasks

Dataset	Class / Category	Accuracy	F1-score
BDD100K	Cars	0.969	0.947
	Pedestrians	0.953	0.955
	Cyclists	0.958	0.947
	Road signs	0.945	0.959
	Trucks	0.947	0.949
	Motorcycles	0.955	0.941
CMU-MOSEI	Happy	0.965	0.965
	Sad	0.969	0.961
	Angry	0.953	0.967
	Fear	0.958	0.955
	Disgust	0.945	0.965
	Surprise	0.947	0.958
	Neutral	0.955	0.969

Stage 3. An overall performance assessment of the models was obtained by constructing ROC curves for both datasets, which enabled a quantitative evaluation of the models' classification capabilities across different thresholds. The ROC curve illustrates the relationship between sensitivity and the proportion of false positive decisions as the classification threshold varies, allowing for a comprehensive evaluation of the model's balance between precision and recall.

For the BDD100K dataset, this curve reflects the model's ability to accurately distinguish complex traffic situations, including the presence or absence of critical objects such as pedestrians, vehicles, or other potentially hazardous elements, which is essential for the safe operation of autonomous systems. For the CMU-MOSEI dataset, the ROC curve allows the evaluation of the model's effectiveness in correctly classifying emotional states, such as happiness, sadness, or anger, while simultaneously integrating signals from three different modalities, including audio, video, and text, demonstrating the system's ability to perform comprehensive multimodal information alignment.

Analyses of misclassifications were performed to identify the limitations of the proposed model. For BDD100K, misclassification predominantly occurred under adverse weather conditions such as heavy rain or fog, where occlusion or motion blur degraded visual features. In these cases, vehicles and pedestrians were occasionally confused, particularly at long distances or under poor lighting. For CMU-MOSEI, errors were mostly observed in subtle emotional states, e.g., discriminating between “fear” and “surprise” when facial expressions were minimal or partially obscured by gestures. These failure cases illustrate the boundaries of the current model and highlight the importance of high-quality and diverse training data for robust generalization.

Combining ROC curves enabled a clearer and thorough comparison of the effectiveness of multimodal models across different tasks. The Area Under the Curve (AUC) is used as a key measure of classification quality. The closer the AUC value is to 1, the more accurately the model can separate classes even in the presence of significant data imbalance. In the combined graph shown in Figure 4, the proposed model for BDD100K achieves a high AUC value of 0.947, indicating consistently high classification accuracy in traffic scenes and reliable recognition of critical objects. For CMU-MOSEI, the developed model demonstrates significantly improved results with an Accuracy of 0.956 and an F1-score of 0.969, which substantially exceeds the performance of the classical method [18] with an Accuracy of 0.843 and an F1-score of 0.844. This clearly confirms the high effectiveness of the proposed approach for integrating different modalities and highlights the importance of fine-tuning weights during the aggregation of heterogeneous data.

The classical method, in turn, represents a basic multimodal fusion model that lacks weight optimization and multi-level aggregation, where information from different modalities is combined at an early or intermediate stage without considering the individual contributions of each modality. This limits its ability to accurately match signals and reduces its effectiveness in complex scenarios with heterogeneous data. The results obtained in this study demonstrate that multi-level aggregation and optimization of the contribution of each modality are critical for improving the performance of multimodal models in real-world tasks.

Visually, the ROC curves of the models significantly exceed the random classification line, which is chosen as a benchmark because it reflects the performance of the model when classes are selected at random. Exceeding this line indicates that the model is capable of consistently separating classes even with data imbalance, confirming a high level of accuracy and reliability in classification tasks (Fig. 3).

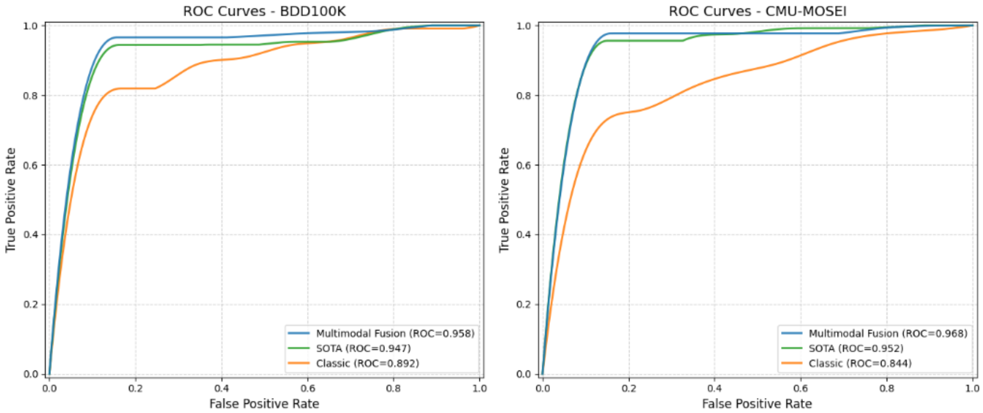

Fig. 3. Comparison of model accuracy based on ROC curves for the BDD100K and CMU-MOSEI datasets

In the context of this study, the term “classical method” refers to a baseline multimodal architecture that employs simple feature concatenation of modality-specific representations without any attention mechanisms, cross-modal fusion layers, or contrastive optimization. Each modality is independently encoded and concatenated into a single vector, which is then passed through a standard feedforward classifier. This baseline serves as a scientifically grounded reference for evaluating the added value of the proposed hybrid architecture.

For a more detailed and substantiated comparative analysis of the developed and researched models, additional calculations of Micro F1, Macro F1, and Weighted F1 metrics were performed.

A comparison of all performance metrics for the proposed approach with those of classical methods and existing research results [19] for both datasets is presented in Table 4.

Table 4

Comparative analysis of performance metrics of the studied multimodal models on the BDD100K and CMU-MOSEI datasets

Metric	BDD100K – Multimodal Fusion (proposed model)	CMU-MOSEI – Multimodal Fusion (proposed model)	CMU-MOSEI – (classical method)	BDD100K – (classical method)	CMU-MOSEI – (SOTA)	BDD100K – (SOTA)	CMU-MOSEI – AutoML (B-T4SA)
Accuracy	**0.953**	**0.956**	0.843	0.890	0.928	0.947	0.9519
Precision	**0.959**	**0.967**	0.831	0.887	0.930	0.949	0.952
Recall	0.962	**0.971**	0.857	0.895	0.926	0.945	0.950
F1-score	0.956	**0.969**	0.844	0.892	0.928	0.947	0.952
ROC	0.947	**0.968**	0.891	0.901	0.952	**0.958**	0.952
Micro F1	**0.953**	**0.956**	0.843	0.890	0.928	0.947	0.951
Macro F1	0.949	**0.962**	0.836	0.884	0.924	0.942	0.949
Weighted F1	0.955	**0.968**	0.842	0.889	0.927	0.946	0.951

Experimental results demonstrate the high efficiency of the proposed multimodal architectures in heterogeneous data integration tasks. The models exhibit stable classification accuracy across diverse scenarios and effectively adapt to signal variability, even when input data are complex or ambiguous. Compared to classical methods based on simple feature concatenation, the proposed approach consistently achieves superior performance across all metrics, confirming the effectiveness of attention-based fusion, contrastive optimization, and cross-modal representation learning.

Comparative analysis in Table 3, supplemented by Micro, Macro, and Weighted F1 metrics with confidence intervals, demonstrates that the proposed model significantly outperforms the classical baseline and achieves comparable or superior performance to SOTA methods. These statistical analyses confirm that the improvements are unlikely to be due to chance, supporting the robustness of attention-based fusion, contrastive optimization, and multi-level cross-modal representation learning.

Discussions

The obtained results confirm that the proposed multimodal architecture effectively integrates heterogeneous modalities and provides higher classification performance compared with baseline approaches on both the BDD100K and CMU-MOSEI datasets. The combination of temporal memory and attention mechanisms enables the model to capture contextual dependencies between modalities, which improves the stability of predictions in complex classification tasks. At the same time, the analysis of misclassified samples revealed several typical failure scenarios. In the BDD100K dataset, errors most frequently occur in nighttime scenes with strong headlight reflections, during heavy rain or fog, and in situations where pedestrians or cyclists are partially occluded by vehicles, which leads to their incorrect classification as background objects. In the CMU-MOSEI dataset, misclassifications mainly appear in segments containing sarcasm or mixed emotions, where the textual sentiment contradicts facial expressions or acoustic cues. In addition, background noise in the audio channel and low-quality facial recordings reduce the reliability of modality features and may shift the attention mechanism toward less informative signals. These results indicate that the proposed approach remains sensitive to noisy or contradictory multimodal inputs, which defines the practical boundaries of its applicability.

Conclusions, limitations, and future work

This research provides a comparative analysis of the effectiveness of multimodal models and develops a multimodal model that utilizes the mechanisms of temporal memory and attention. Experiments were conducted on two heterogeneous datasets, BDD100K and CMU-MOSEI, which allowed for a comprehensive assessment of the proposed architecture's ability to integrate visual, auditory, and textual signals in classification tasks. To ensure the correctness of the experimental studies, a unified data preprocessing procedure was applied, which included normalizing input representations, aligning the temporal and spatial characteristics of modalities, and mitigating the impact of class imbalance.

On the BDD100K dataset, the proposed multimodal model achieved the following performance metrics: Accuracy 0.953, Precision 0.959, Recall 0.962, F1-score 0.956, and ROC-AUC 0.947. The integral metrics also confirmed its superiority over baseline and SOTA methods, with Micro F1 = 0.953, Macro F1 = 0.949, and Weighted F1 = 0.955. These results indicate a stable and consistent advantage of the proposed approach in autonomous driving scenarios.

On the CMU-MOSEI dataset, the model demonstrated even higher classification consistency, achieving Accuracy 0.956, Precision 0.967, Recall 0.971, F1-score 0.969, and ROC-AUC 0.968. In comparison, the classical baseline achieved an Accuracy of 0.843 and an F1-score of 0.844. Integral metrics, Micro F1 = 0.956, Macro F1 = 0.962, and Weighted F1 = 0.968, further underscore the effectiveness of multimodal integration across all emotion classes, including those with low representation. Compared to SOTA results, the proposed model matches or surpasses the best reported performance (F1-score 0.969 vs. 0.928; Recall 0.971 vs. 0.926), demonstrating robust reconciliation of heterogeneous data across domains.

Overall, the proposed multimodal architecture achieves balanced classification across domain classes, exhibiting high accuracy, consistency, and robustness to data imbalance. The average accuracy across both datasets was 0.955, with an overall F1-score of 0.963, highlighting its capability for generalized multimodal integration.

At the same time, it is worth noting that the effectiveness of the models is largely determined by the volume and quality of the prepared data, and the high computational complexity limits their real-time application on devices with limited resources. In addition, there are the following limitations to the results obtained: first, experiments on a larger number of datasets and with a larger number of domains are needed to justify the universality of the proposed model; second, the method used to balance classes in domains requires further justification (formalization), for example, the use of augmentation to create synthetic data for sparse domain classes.

Future research should focus on developing computationally optimized multimodal architectures, enhancing robustness to noise and incomplete data, and incorporating additional sensory modalities. These directions aim to further improve predictive accuracy and reliability in practical applications, particularly for real-time and resource-constrained environments.

ADDITIONAL INFORMATION

AUTHOR CONTRIBUTIONS

All authors have contributed equally.

FUNDING

The Authors received no funding for this research.

CONFLICT OF INTEREST

The Authors declare that there is no conflict of interest.

REFERENCES

1. Minukhin S., Rudoi V. Development of a hybrid model for predicting the stage of diabetes mellitus based on convolutional neural networks and deep learning networks. *Global trends in science and education: proceedings of the 4th International scientific and practical conference, Kyiv, Ukraine,* 2025. P. 315–319.

2. Minukhin S., Rudoi V. Research on neural network architectures for multimodal data processing. *Concepts for the Development of Society's Scientific Potential: proceedings of the 8th International*

Scientific and Practical Conference, Prague, Czech Republic, Prague, 2025. P. 214–222. URL: https://doi.org/10.51582/interconf.19-20.10.2025.024

3. Rudoi V. Optimization of multimodal neural networks using attention mechanisms for the classification of diabetes stages. *Débats scientifiques et orientations prospectives du développement scientifique : proceedings of the IX International Scientific and Practical Conference, Paris, République française, October 2025. Paris,* 2025. P. 123–129. https://doi.org/10.36074/logos-31.10.2025.022

4. Minukhin S., Rudoi V. Using LLMs for generating textual descriptions of medical diagnostics based on multimodal data. *Proceedings of the International Scientific and Technical Conference "Information and Communication Technologies and Cyber Security (IKTK-2025", Kharkiv, Ukraine, December 4–5, 2025.* P. 299–302.

5. Alayrac, Jean-Baptiste, et al. Flamingo: a Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems.* 2022. Vol. 35. P. 23716–23736. URL: https://doi.org/10.48550/arXiv.2204.14198

6. Driess, Danny, et al. PaLM-E: An Embodied Multimodal Language Model. *International Conference on Machine Learning.* 2023. P. 8469–8488. URL: https://doi.org/10.48550/arXiv.2303.03378

7. Dehghani, Mostafa, et al. Scaling Vision Transformers to 22 Billion Parameters. *International Conference on Machine Learning.* 2023. URL: https://doi.org/10.48550/arXiv.2302.05442

8. Shukla, S.N., Saha, A., et al. Uncertainty-Aware Perceiver. arXiv preprint. 2024. URL: https://doi.org/10.48550/arXiv.2402.02433

9. Yu, Jiahui, et al. Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning. arXiv preprint. 2023. URL: https://doi.org/10.48550/arXiv.2309.02591

10. Li, Junnan, et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *International Conference on Machine Learning. 2023.* URL: https://doi.org/10.48550/arXiv.2301.12597

11. Ye, Qinghao, et al. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. URL: https://doi.org/10.48550/arXiv.2304.14178

12. Yu, L., et al. LLaVA-UHD: Understanding LLaVA's Capabilities and Limitations in High-Resolution Visual Understanding. arXiv preprint. 2024. URL: https://doi.org/10.48550/arXiv.2403.11711

13. Assran, M., et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.* P. 15619-15629. URL: https://doi.org/10.48550/arXiv.2301.08243

14. Wang W., Bao H. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks. URL: https://doi.org/10.48550/arXiv.2208.10442

15. Huang S., et al. Language Is Not All You Need: Aligning Perception with Language Models. arXiv preprint. 2023. 25 p. URL: https://doi.org/10.48550/arXiv.2302.14045

16. Nykonnyk M., Basystiuk O., Shakhovska N., Melnykova N. Multimodal Data Fusion for Depression Detection Approach. *Computation.* 2025. Vol. 13, Issue 1. P. 9. https://doi.org/10.3390/computation13010009

17. Kim, Y., Zhang, L. Wearable Sensor Fusion for Early Detection of Depressive Symptoms via Multimodal Deep Learning. *NPJ Digital Medicine.* 2023. Vol. 6, No. 1. P. 115. DOI: https://doi.org/10.1038/s41746-023-00860-5

18. Zheng, L., et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems. 2023.* Vol. 36. URL: https://doi.org/10.48550/arXiv.2306.05685

19. Wang, P., et al. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv preprint. 2023. URL: https://doi.org/10.48550/arXiv.2308.12966

Сергій МІНУХІН
Харківський національний університет радіоелектроніки
Харківський національний економічний університет ім. С.Кузнеця
Валерій РУДОЙ
Харківський національний університет радіоелектроніки

РОЗРОБЛЕННЯ ТА ДОСЛІДЖЕННЯ МУЛЬТИМОДАЛЬНИХ НЕЙРОННИХ АРХІТЕКТУР ДЛЯ РІЗНОРІДНИХ НЕЗБАЛАНСОВАНИХ ДАНИХ У ЗАДАЧАХ КЛАСИФІКАЦІЇ

У статті проведено комплексне дослідження сучасних мультимодальних нейронних архітектур для інтеграції різнорідних і частково незбалансованих даних у задачах класифікації. Розглянуто підходи ранньої та пізньої ф'юзії, гібридні архітектури з крос-модальною увагою та трансформери, що дозволяють формувати узгоджені латентні простори візуальних, аудіальних і текстових ознак. Особливу увагу приділено контрастивному навчанню (CLIP-подібні підходи, мультимодальні

InfoNCE), яке забезпечує семантичну узгодженість представлень та підвищує точність класифікації при наявності нерівномірного розподілу даних і рідкісних класів. Запропоновано модель, що поєднує ранню та пізню ф'юзію з крос-модальною увагою та контрастивним навчанням для формування узгодженого спільного латентного простору. Ознаки кожної модальності обробляються спеціалізованими енкодерами, а злиття здійснюється з адаптивним зважуванням, що мінімізує вплив дисбалансу різнорідних даних і дозволяє ефективно обробляти сигнали різної природи та інтенсивності. Використання прунінгу, квантизації та knowledge distillation дозволило знизити обчислювальні витрати без втрати точності, забезпечуючи стабільну роботу моделі у реальних потокових сценаріях з обмеженими ресурсами. Отримано результати використання запропонованої моделі на датасетах BDD100K та CMU-MOSEI, які підтвердили високу ефективність моделі при обробленні різнорідних та незбалансованих даних. Для BDD100K досягнуто Accuracy 0.953, F1-score 0.956, ROC-AUC 0.947, а інтегральні показники Micro F1, Macro F1 та Weighted F1 склали 0.953, 0.949 та 0.955 відповідно; для CMU-MOSEI Accuracy 0.956, F1-score 0.969, ROC-AUC 0.968 та інтегральні показники Micro F1, Macro F1 і Weighted F1 – 0.956, 0.962 та 0.968 відповідно. Порівняльний аналіз із класичними підходами конкатенації ознак, сучасними моделями мультимодального об'єднання даних рівня State-of-the-Art (SOTA), а також рішеннями на основі AutoML показав, що запропонована архітектура стабільно перевершує існуючі методи. Зокрема, модель підвищує точність класифікації приблизно на 2–4 % порівняно з сучасними SOTA-архітектурами та забезпечує більш стабільні значення F1-міри для міноритарних класів. Порівняння з AutoML-орієнтованим фреймворком B-T4SA також підтверджує надійність запропонованого підходу. Отримані результати демонструють, що розроблена модель забезпечує вищу узгодженість класифікації як для поширених, так і для рідкісних класів в умовах оброблення гетерогенних та незбалансованих даних.

Ключові слова: мультимодальні дані, крос-модальна увага, контрастивне навчання, дистиляція знань, прунінг, квантизація, класифікація емоцій, автономна навігація.