Multimodal Data Integration for Smart Industrial Automation Pipelines*

Andrii Dumyn[1,†] and Oleh Basystiuk[2,†]

[1] N-iX, 157 Archbishop Street, Valletta, VLT 1440, Malta
[2] New Western University, 970 S Village Oaks Dr. Suite 212, Covina, CA 91724, USA

Abstract

Modern industrial enterprises require flexible and intelligent quality control systems that are able to analyze different types of data simultaneously. This paper considers the use of multimodal data to automate defect detection, predict equipment failures, and improve the efficiency of production processes. The main focus is on combining text, audio, and image information to create a single analytical system. For example, analyzing video from a conveyor in combination with audio vibration diagnostics allows you to detect deviations that cannot be captured using separate sensors. Considered using multimodal data to automate defect detection, predicting equipment failures and monitoring the emotional state of the employee. The main focus is on combining text like HMI logs, audio and imaging information to create a single analytical system. The paper proposes a late and adaptive weighted data fusion based on transformers and deep neural networks, which allows to increase the accuracy control and reduce the probability of false equipment responses. Experiments would be conducted on real production data using modern ML frameworks.

Keywords

multimodal data integration, human-centeredness, cross-modal attention, adaptive weighted fusion, industrial anomaly detection, predictive maintenance, human-machine interface

1. Introduction

The evolution of industrial production has gone from the mechanization of the first revolution to the digital integration of Industry 4.0, where automation, cyber-physical systems and data-driven efficiency were the key focuses. However, the current global context requires a transition to Industry 5.0, a paradigm that does not replace but complements the previous one, placing human-centricity, sustainability and environmental sustainability at the heart of the production process. In this new environment, intelligent automation must not only replace human labor in routine operations, but also enhance the cognitive capabilities of workers, ensuring safety and health through the deep integration of artificial intelligence technologies.

For the analysis of complex industrial processes, the use of individual sensors is often insufficient due to the high level of noise and the complexity of the relationships between parameters. Multimodal models that combine video data from conveyors, acoustic diagnostics of engine vibrations and analysis of operator log files allow you to detect hidden patterns and correlations. For example, changes in the spectral characteristics of the sound of equipment can signal the degradation of bearings much earlier than the defect becomes visually noticeable. In addition, the integration of human-machine interfaces and personnel reports allows for the consideration of "soft" data that is often ignored by traditional monitoring systems.

The relevance of this research is reinforced by the need to create fault-tolerant systems capable of operating in conditions of uncertainty, which is critically important for Ukrainian enterprises during the period of martial law and subsequent reconstruction [1]. The implementation of

* SmartIndustry 2026: 3rd International Conference on Smart Automation & Robotics for Future Industry, March 26-27, 2026, Lviv, Ukraine
1* Corresponding author.
† These authors contributed equally.
andrew.dmn@gmail.com (A. Dumyn); obasystiuk@nwuus.org (O. Basystiuk);
0000-0003-2111-2899 (A. Dumyn); 0000-0003-0064-6584 (O. Basystiuk)

Industry 5.0 technologies allows us to adapt to labor shortages and infrastructure destruction through the use of autonomous agents and digital twins.

2. Analysis of architectural features of multimodal systems

Developing an effective data integration system requires understanding the specifics of each modality and the methods for combining them. In the industrial context, there are three dominant data types, each of which requires a specific approach to preprocessing and feature extraction [2].

Images and video streams from quality control cameras are the main source for detecting surface defects: cracks, dents, chips or violations of the integrity of connections. Modern approaches are based on convolutional neural networks, such as ResNet or EfficientNet, which provide high classification accuracy at moderate computational costs. For detailed localization of defects, segmentation based on U-Net-type architectures is used, which allows to highlight anomalous areas at the pixel level [3, 4, 5].

The evolution of visual inspection methods has led to the creation of systems capable of unsupervised learning based on normal samples alone. This is especially important for industries where defect samples are rare and variable [6].

The sound of the equipment is the "voice" of the machine, carrying information about internal processes. The use of microphones and vibration sensors allows you to record anomalies that are inaccessible to vision. The main processing method is the conversion of sound signals into the spectral domain using short-time Fourier transform or the calculation of Mel-spectrograms and MFCC.

Recurrent neural networks or transformers are used to analyze time dependencies in sound patterns, allowing the identification of beats, non-standard noises, or changes in frequency characteristics.

Textual information includes control system log files, technical documentation, and operator reports. Using natural language processing models such as BERT or T5, it is possible to extract semantic context from error messages. Analysis of the tone of operator reports can complement automatic diagnostics, for example, when an employee indicates unusual equipment behavior that has not yet triggered critical sensor thresholds [7].

3. Proposed approach

The central problem of multimodal learning is the efficient fusion of heterogeneous features. In this paper, we propose an adaptive weighted fusion architecture based on attention mechanisms.
Let us consider a set of input data for different modalities:

x_{img} – sequence of video frames or photos;

x_{audio} – acoustic signals;

x_{text} – text reports and logs.

The challenge is to transform this data into a shared latent space.

3.1. Feature extraction by modalities

For each modality, a specialized neural network architecture is used, optimized for the corresponding data type.

Visual data is processed using convolutional neural networks such as ResNet or EfficientNet [8]. Image features are formed as:

$$F_{img} = CNN(x_{img}, \theta_{img}), \quad (1)$$

where x_{img} is the input image, and θ_{img} is the trained network parameters.

Audio signals are first converted into mel-spectrograms:

$$S = \text{Mel\_Spectrogram}(\text{STFT}(x_{audio})), \quad (2)$$

which allows you to convert a time signal into a time-frequency representation.

Next, features are extracted using recurrent or transformer models:

$$F_{audio} = g_{\phi}(S), \quad (3)$$

where $g_{\phi}$is an LSTM, BiLSTM, or Transformer encoder with parameters ϕ.

Text information is encoded as embedding vectors using transform language models such as BERT or DeBERTa:

$$F_{text} = t_{\psi}(x_{text}), \quad (4)$$

where $x_{text}$is the text input, and $t_{\psi}$is the text encoder with parameters ψ.

Received modal representations F_{img}, F_{audio}, F_{text} are further used for multimodal fusion and formation of a common latent feature space.

Each modality $m \in \{img, audio, text\}$, processed by a specialized encoder E_m, which forms a feature vector $f_m \in R^{d_m}$.

Since the latent representations of different modalities have different dimensions and statistical properties, before the multimodal fusion stage, their projection into a common latent space of dimension is performed. d.

The projection is implemented through a nonlinear transformation:

$$\hat{f}_m = \phi(f_m, W_p) = \sigma\left(W_p^{(m)} f_m + b_p^{(m)}\right), \quad (5)$$

where $W_p^{(m)} \in R^{d \times d_m}$ – projection matrix for modality m; $b_p^{(m)}$ – displacement vector; $\sigma(\cdot)$ – nonlinear activation function.

Such a transformation ensures the consistency of the dimensions of modal features and allows the integration of heterogeneous representations into a common multimodal space for further fusion analysis.

To manage the high heterogeneity of industrial indicators (visual defects, telemetry and reports), it is advisable to adapt the graph document-oriented structure, originally developed for medical systems. This approach allows you to transform disparate data streams into a coherent graph template, which provides effective modeling of complex nonlinear relationships between the technical condition of the equipment and external factors within a single analytical pipeline [9].

An important aspect of industrial automation reliability is the processing of semi-structured records (system logs, telemetry, and events), where concepts of efficient medical data storage can be used as an architectural prototype. The implementation of such models allows for optimizing the aggregation of multimodal time series, providing fast access to historical data for accurate prediction of residual life and minimizing false alarms [10].

3.2. Adaptive weighted fusion

Since different modalities have different informativeness depending on the context of the task, the system uses the adaptive weight fusion mechanism [12]. For this purpose, trainable importance coefficients are introduced w_i, which are dynamically adjusted by the attention mechanism [13]. The integrated multimodal representation is formed as:

$$F_{fusion} = \sum_{i=1}^{M} w_i \cdot \phi(F_i), \quad (6)$$

where F_i – signs i of the -th modality; $\phi(\cdot)$ – nonlinear transformation for matching the dimensions of latent spaces; w_i — weights of modalities; $\sum_i w_i = 1$– ensuring normalization.

Weights can either be parameters learned directly or calculated dynamically via a Gating Network:

$$w_m = \frac{\exp\left(g_m^T \hat{f}_m\right)}{\sum_j \exp\left(g_j^T \hat{f}_j\right)}, \tag{7}$$

where g_m — attention vector for modality m.

This mechanism allows the model to automatically change the degree of confidence in individual modalities depending on the signal quality and the context of the task.
The scheme of adaptive feature fusion is shown in Figure 1.

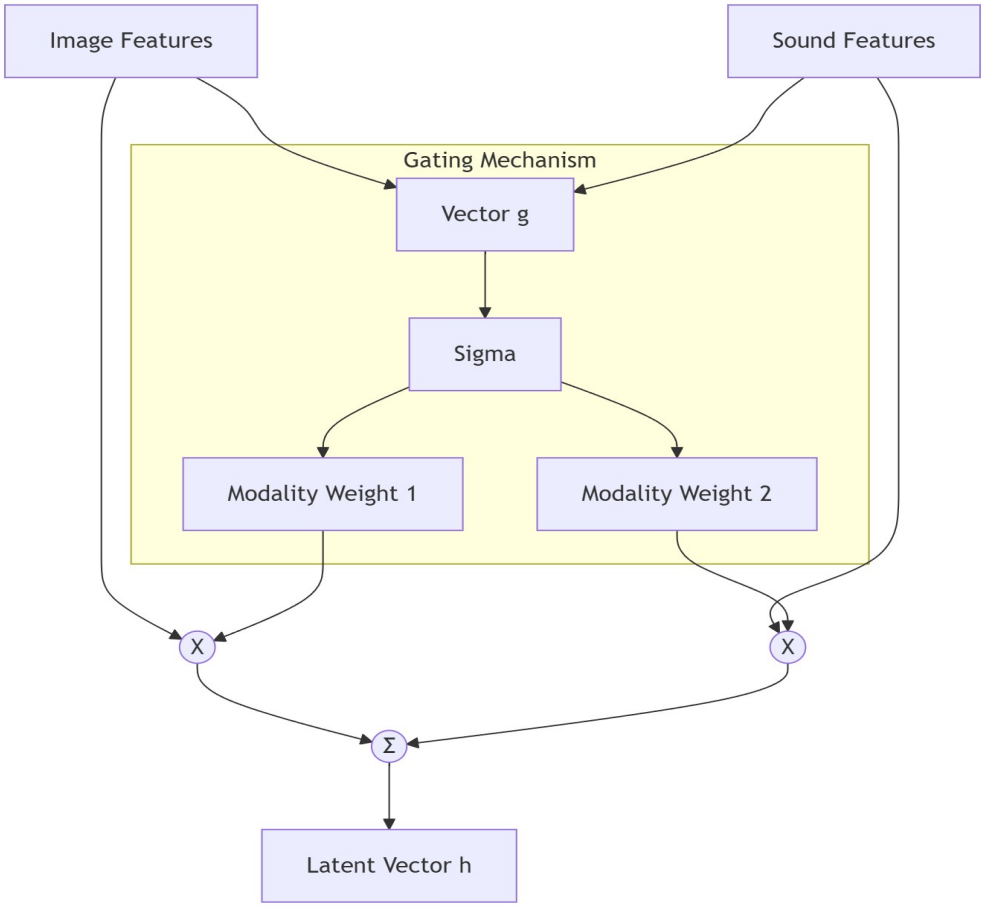

Figure 1: Scheme of adaptive feature fusion.

To model deep interaction between different modalities, in particular text and visual, the cross-attention mechanism is used [14]. Let Q_{text} – the query matrix (Query) from the text modality; K_{img} – the key matrix (Key) from the visual modality; V_{img} – the value matrix (Value) of the image.

Then the attention mechanism is defined as:

$$A\left(Q_{text}, K_{img}, V_{img}\right) = \text{softmax}\left(\frac{Q_{text} K_{img}^T}{\sqrt{d_k}}\right) V_{img}, \tag{8}$$

where d_k is the dimension of the key space.

The logic of Cross-Attention Fusion is shown in Figure 2.

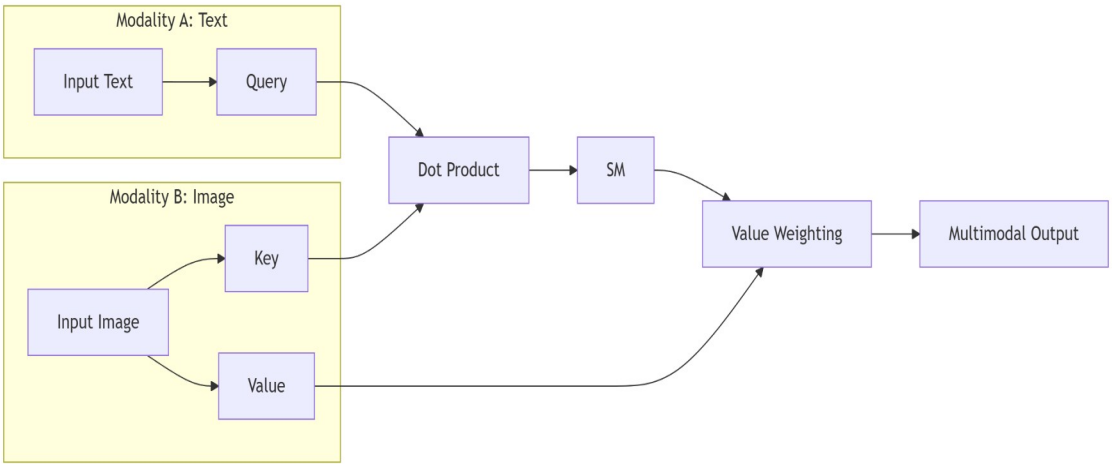

Figure 2: Conceptual diagram of the proposed approach.

This allows the system to adaptively focus on the most relevant cross-modal dependencies. For example, the model can increase attention to certain visual features when a characteristic audio signal appears, or, conversely, use textual context to refine the interpretation of audio or video data.

Figure 3 shows the architecture of the proposed system.

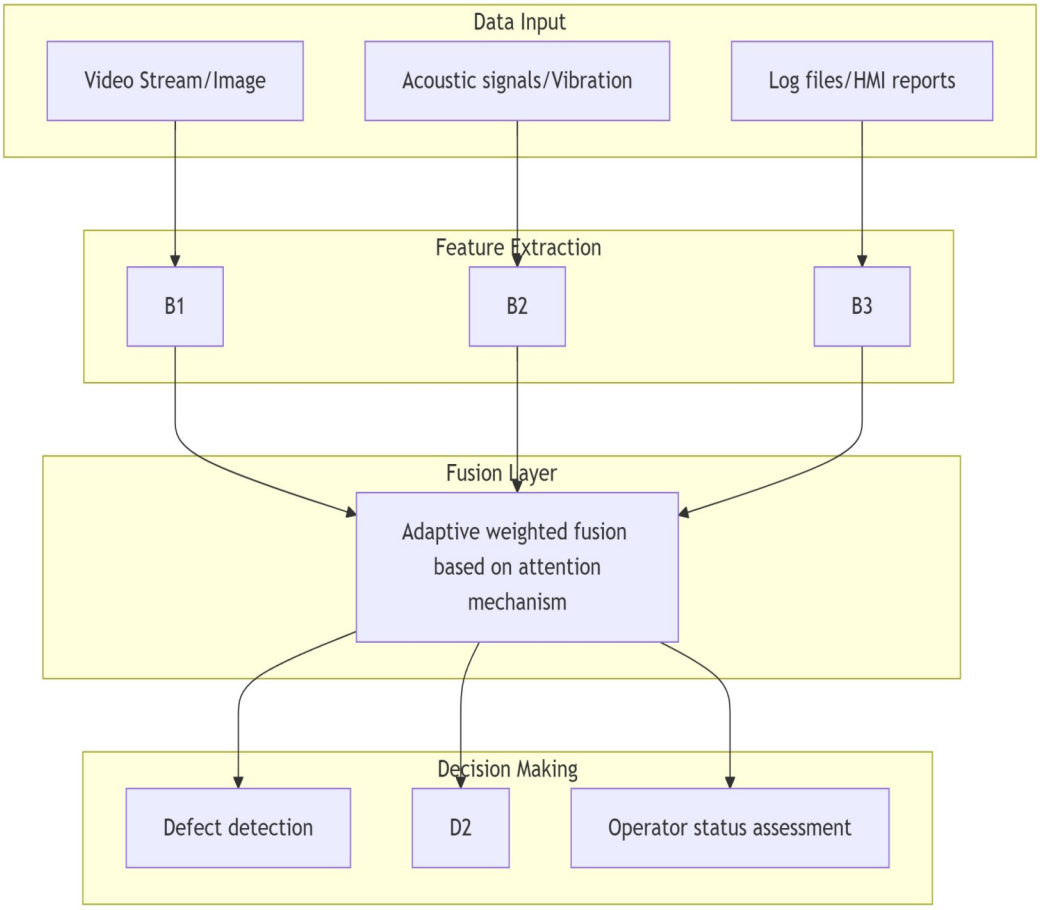

Figure 3: Architecture diagram of a multimodal industrial automation.

3.3. Classification and optimization

Merged multimodal vector *F*, is fed to a fully connected classification layer:

$$\hat{y} = \sigma(W_{class} F + b), \tag{9}$$

where W_{class} and b – classifier parameters; $\sigma(\cdot)$ – activation function.

For multi-class defect classification problems, the Softmax function is used:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}, \tag{10}$$

while for binary anomaly detection Sigmoid is used:

$$\sigma(z) = \frac{1}{1+e^{-z}}. \tag{11}$$

Model training is performed using combined loss functions depending on the type of problem.

For classification tasks, the cross-entropy loss function is used:

$$L_{CE} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c), \tag{12}$$

where y_c – true class label; $\hat{y}_c$ – predicted class probability.

For regression problems predicting the remaining life of equipment, the root mean square or mean absolute error functions are used:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{13}$$

or

$$L_{MAE} = \frac{1}{N}\sum_{i=1}^{N} /y_i - \hat{y}_i /. \tag{14}$$

Model parameters are optimized using the Adam algorithm with adaptive learning rate. η_t, which allows to stabilize the training process and accelerate convergence in multimodal architectures.

4. Industry 5.0 and emotional intelligence

One of the key innovations of the proposed system is the focus on the human factor in accordance with the concept of Industry 5.0, within which the employee is considered not as a source of production variability, but as an active participant in human-machine interaction. Unlike Industry 4.0 approaches, where the priority was maximum automation, Industry 5.0 focuses on the well-being, safety and cognitive support of personnel.

Figure 4 demonstrates how worker health data is integrated into production process management to ensure safety and productivity.

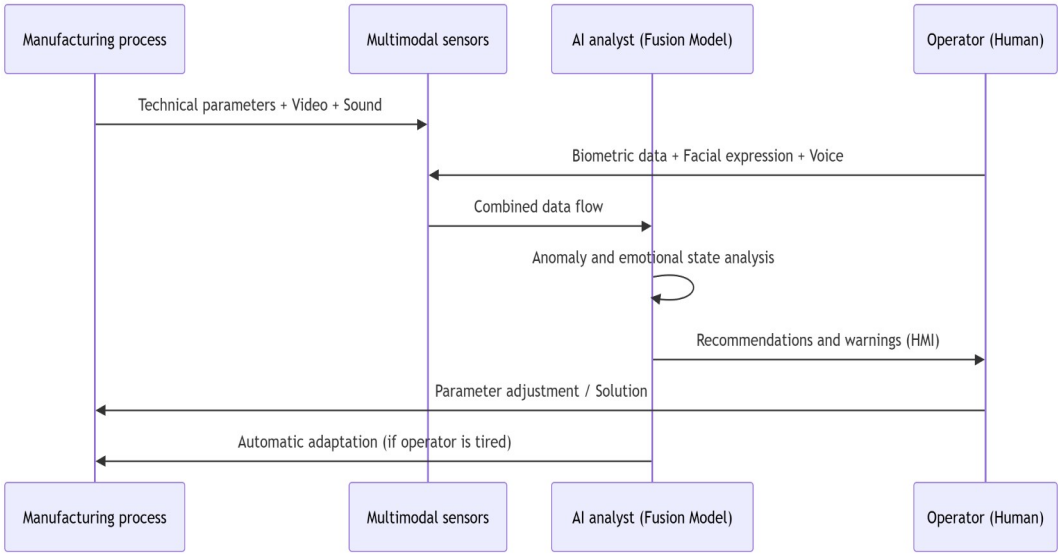

Figure 4: Closed loop of human-machine interaction.

Multimodal artificial intelligence opens the possibility of monitoring the psycho-emotional state of operators in real time by integrating several types of data. Facial expression analysis allows you to detect signs of fatigue, stress or anxiety, acoustic voice analysis – to assess cognitive overload through changes in tone and pace of speech, and text analysis of interaction with HMI systems – to determine the level of concentration of attention based on the speed of command entry and the nature of reports [15]. Additionally, physiological sensors, in particular infrared temperature and breathing sensors, can be used to assess the functional state of the employee.

The integration of the Human Digital Twin concept allows for the creation of a digital model of the worker's state and the adaptation of the production environment to their current physical and emotional workload [16]. In practice, this may include personalized scheduling, automatic adjustment of lighting, conveyor speed, or the intensity of production processes. This approach not only improves safety, but also reduces burnout, employee turnover, and healthcare costs.

The implementation of the Industry 5.0 paradigm requires the improvement of human-machine interfaces by integrating emotional voice control. The use of multimodal signals to classify the emotional state of the operator allows creating an adaptive control system for industrial IoT devices, which can automatically adjust the parameters of the production process depending on the level of cognitive load or fatigue of personnel [17].

At the same time, the implementation of such systems raises serious ethical and social challenges. Constant monitoring of emotional state can create risks of excessive digital control, loss of privacy and the formation of the effect of "hyper-surveillance". In this regard, modern research emphasizes the need to ensure the transparency of algorithms, explainability of decisions and strict adherence to the principles of personal data protection. In the context of the European model of Industry 5.0, cybersecurity, ethical use of AI and preservation of human autonomy are considered as fundamental components of human-centric industrial systems.

5. Conclusions and discussion

The analysis demonstrates that the future of industrial automation lies in the deep integration of multimodal data. The combination of visual, acoustic and text sources allow overcoming the limitations of traditional monitoring systems, providing accuracy that was previously unavailable.

The use of adaptive fusion mechanisms based on transformers allows models to be resistant to noise and dynamically change the priority of information sources.

However, technological progress must be accompanied by a humanitarian component. Industry 5.0 proposes a paradigm shift where AI becomes a partner of humans, taking care of their safety and emotional state. This creates new opportunities for increasing work efficiency while simultaneously improving the quality of life of workers.

For Ukraine, the introduction of these technologies is a chance for a qualitative leap in industrial development. The creation of intelligent conveyors based on the principles of human-centricity and sustainability will allow not only to rebuild what has been destroyed, but also to lay the foundation for a modern, high-tech industrial ecosystem integrated into the global space.

Further research in the field of multimodal artificial intelligence for industrial systems will be directed towards several key directions. First, the development of lightweight multimodal architectures suitable for operation on edge devices in real time remains relevant, which is critically important for Industrial IoT and cyber-physical systems. An equally promising direction is the improvement of reinforcement learning methods for adaptive optimization of production parameters based on multimodal feedback from sensors, operators and the production environment.

Particular attention should be paid to the formation of ethical and legal frameworks for the use of emotional AI in industrial settings, in particular regarding the transparency of algorithms, the protection of personal data, and the prevention of discriminatory practices. This is especially important in the context of the transition to human-centric models of Industry 5.0.

In general, multimodal data integration serves not only as a technological tool for increasing production efficiency, but also as the foundation of a new paradigm of intelligent production, in which artificial intelligence is focused on supporting humans, increasing safety, and ensuring the sustainable development of society.

Declaration on Generative AI

During the preparation of this work, the author used AI tools in order to: Grammar and spelling check.

References

[1] View of Industry 5.0 as a Tool to Ensure the Effective Development of Ukrainian Enterprises During Military Challenges, https://reicst.com.ua/pmt/article/view/2024-12-03-07/2024-12-03-07

[2] Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A Survey on Deep Learning for Multimodal Data Fusion. Neural Computation, 32, 829-864. https://doi.org/10.1162/neco_a_01273 .

[3] Kullu, O., & Cinar, E. (2022). A Deep-Learning-Based Multi-Modal Sensor Fusion Approach for Detection of Equipment Faults. Machines. https://doi.org/10.3390/machines10111105 .

[4] Rahman, M., Hossain, M., Rozario, U., Roy, S., Mridha, M., & Dey, N. (2025). MultiSenseNet: Multi-Modal Deep Learning for Machine Failure Risk Prediction. IEEE Access, 13 , 120404-120416. https://doi.org/10.1109/access.2025.3586978 .

[5] Yang, Z., Baraldi, P., & Zio, E. (2021). A multi-branch deep neural network model for failure prognostics based on multimodal data. Journal of Manufacturing Systems, 59 , 42-50. https://doi.org/10.1016/j.jmsy.2021.01.007 .

[6] Zhao, Y., Zhang, Y., Li, Z., Bu, L., & Han, S. (2023). AI-enabled and multimodal data driven smart health monitoring of wind power systems: A case study. Adv. Eng. Informatics, 56 , 102018. https://doi.org/10.1016/j.aei.2023.102018 .

[7] McKinney, M., Garland, A., Cillessen, D., Adamczyk, J., Bolintineanu, D., Heiden, M., Fowler, E., & Boyce, B. (2025). Unsupervised multimodal fusion of in-process sensor data for advanced manufacturing process monitoring. Journal of Manufacturing Systems . https://doi.org/10.1016/j.jmsy.2024.12.003 .

[8] Tamakloe, E., Kommey, B., Kponyo, J., Tchao, E., Agbemenu, A., & Klogo, G. (2025). Predictive AI Maintenance of Distribution Oil-Immersed Transformer via Multimodal Data Fusion: A New Dynamic Multiscale Attention CNN-LSTM Anomaly Detection Model for Industrial Energy Management. IET Electric Power Applications . https://doi.org/10.1049/elp2.70011 .

[9] I. Dumyn, O. Basystiuk, and A. Dumyn, “Graph-based approaches for multimodal medical data processing,” in Proc. 7th Int. Conf. on Informatics & Data-Driven Medicine, Birmingham, UK, Nov. 2024, pp. 349–362.

[10] I. Shvorob, “New approach for saving semistructured medical data,” in Advances in Intelligent Systems and Computing, vol. 512, N. Shakhovska, Ed. Cham, Switzerland: Springer, 2017, pp. 29–39. doi: 10.1007/978-3-319-45991-2_3.

[11] Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A Survey on Deep Learning for Multimodal Data Fusion. Neural Computation, 32 , 829-864. https://doi.org/10.1162/neco_a_01273 .

[12] Zhao, F., Zhang, C., & Geng, B. (2024). Deep Multimodal Data Fusion. ACM Computing Surveys, 56 , 1 - 36. https://doi.org/10.1145/3649447 .

[13] Sui, W., Lichau, D., Lefèvre, J., & Phelippeau, H. (2024). Incomplete multimodal industrial anomaly detection via cross-modal distillation. Inf. Fusion, 126 , 103572. https://doi.org/10.1016/j.inffus.2025.103572 .

[14] Villegas-Ch., W., Gaibor-Naranjo, W., & Sánchez-Viteri, S. (2024). Application of Deep Learning Techniques for the Optimization of Industrial Processes Through the Fusion of Sensory Data. International Journal of Computational Intelligence Systems, 17. https://doi.org/10.1007/s44196-024-00596-4 .

[15] Paikrao, P., Mukherjee, A., Guled, C., Goswami, P., & Narwade, P. (2025). Smart Manufacturing in Industrial AIoT 5.0 Applications: A Speech Emotion Recognition Approach. IEEE Internet of Things Journal, 12 , 42693-42701. https://doi.org/10.1109/jiot.2025.3594566 .

[16] Kedanjoth, C., Thomas, M., Pohren, D., Roque, A., & Freitas, E. (2025). An IoT-based Multimodal AI System for Emotional and Behavioral Analysis. 2025 12th International Conference on Future Internet of Things and Cloud (FiCloud) , 151-158. https://doi.org/10.1109/ficloud66139.2025.00029 .

[17] Shevchuk I., Dumyn I. Emotion-based voice control for IoT: enhancing smart device interaction with speech emotion classification // CEUR Workshop Proceedings. - 2025. - Vol. 3974 : Joint proceedings of the workshops "AI for environmental and social sustainability workshop" and "AI and interdisciplinary innovations for sustainable development" (YAISD-WS 2025) co-located with Second international conference of young scientists on artificial intelligence for sustainable development (YAISD 2025) Ternopil-Skomorochy, May 8-9, 2025. – P. 196–203.