Multimodal Sensor Data Fusion for Robust Deep Learning Models in Manufacturing Applications*

Zoriana Rybchak[1,†] and Oleh Basystiuk[2,†]

[1] Researcher at State-Funded Project "Multisensor", 15 Mytropolyta Andreia St., Building 4, Room 122a, 79000, Lviv, Ukraine
[2] New Western University, 970 S Village Oaks Dr. Suite 212, Covina, CA 91724, USA

Abstract

This paper considers the integration of multimodal sensor data fusion to improve the reliability of deep learning models in smart manufacturing. Methods for collecting, storing, and processing heterogeneous data are considered, as are neural network architectures that ensure fault tolerance and decision explainability. The development of the Industry 4.0 concept is impossible without deep integration of operational technologies with information systems, with the Industrial Internet of Things playing a central role. A modern shop floor generates vast amounts of big data, including log files, signal streams, and telemetry. For effective analysis of this data, real-time systems are needed to detect patterns in large historical datasets. To effectively process these diverse modalities, this paper proposes a late and adaptive weighted data fusion architecture based on transformers and deep neural networks. By combining text, audio, and imaging information into a single analytical framework, the proposed system significantly improves diagnostic accuracy, ensures fault tolerance, and reduces the probability of false equipment responses. The efficacy of the proposed models will be evaluated through experiments conducted on real production data using modern machine learning frameworks.

Keywords

multimodal data , sensors, smart manufacturing, data fusion, deep learning, Industry 4.0, Internet of Things, cross-modal attention, robust models, human-machine interface

1.Introduction

The evolution of industrial production has transitioned from the mechanization of early industrial revolutions to the digital integration of Industry 4.0, where automation, cyber-physical systems, and data-driven efficiency take center stage. In modern manufacturing, the integration of multimodal data is emerging as the foundation for creating intelligent systems capable of autonomous condition monitoring and diagnostics.

In modern industry, the integration of multimodal data is becoming the foundation for creating systems that are able to independently diagnose their own condition. The work focuses on how combining information from different types of sensors allows us to increase the accuracy of neural network predictions. We consider collection methods, storage methods in heterogeneous environments, as well as network architectures that ensure production resilience to failures. The development of the Industry 4.0 concept is impossible without a deep combination of operational technologies with information systems, where the Industrial Internet of Things plays a central role. A modern shop floor generates huge arrays of big data, which include log files, signal streams, and telemetry. To effectively analyze this data, real-time systems are needed that are able to detect patterns in huge historical array

The development of the Industry 4.0 [1] concept depends entirely on the seamless combination of operational technologies and information systems. Modern manufacturing facilities generate massive arrays of historical and real-time big data. This paper explores data collection and heterogeneous storage methods, alongside robust deep learning network architectures designed to ensure production resilience against sensor failures and environmental uncertainty. By compensating for the limitations of individual modalities, multimodal fusion provides a robust, fault-tolerant framework critical for the continuous operation of modern enterprises.

*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.
Zoriana.rybchak@gmail.com (Z. Rybchak) ; obasystiuk@nwuus.org (O. Basystiuk);
0000-0002-5986-4618 (Z. Rybchak) ; 0000-0003-0064-6584 (O. Basystiuk)

The Figure 1 illustrates the conceptual structure of multimodal distribution analysis. The innermost circle (yellow) represents Mode 1 — the highest peak in the distribution, which identifies the dominant trend or the most frequently occurring cluster of values in the dataset. Surrounding it, the second circle (orange) denotes Mode 2 — a secondary peak that indicates the presence of an alternative trend or a subpopulation within the data, characteristic of bimodal or multimodal behavior. The third ring (red/pink) — the low-frequency regions between peaks that reflect data gaps and serve as natural boundaries separating distinct modal clusters. The outermost circle (green) encompasses Clustering Patterns — the broadest level of analysis, capturing the overall grouping of data points around the identified modes and revealing the general distributional structure of the dataset.

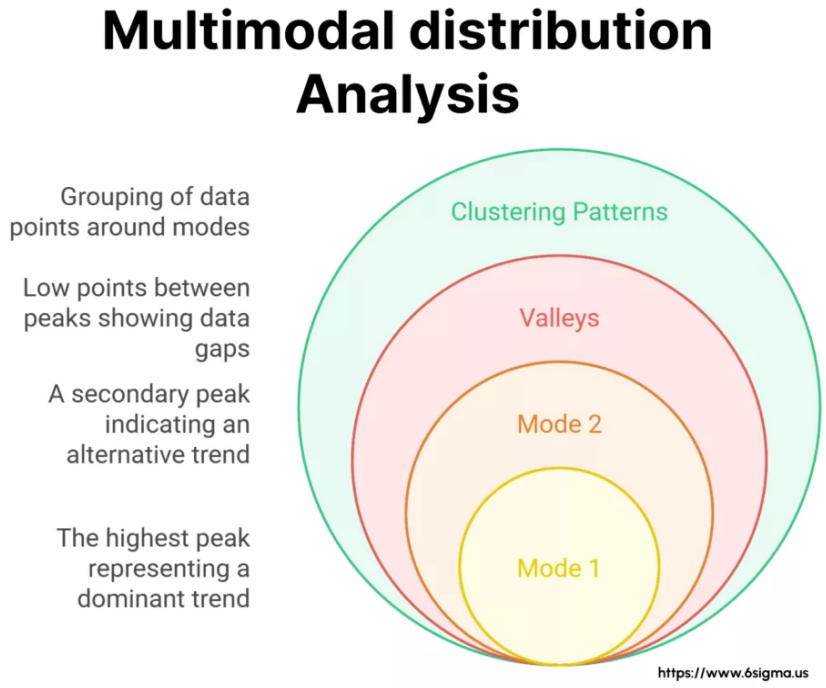

Figure 1: Structure of multimodal data.

2.Analysis of integration approach with Internet of Things systems

The development of Industry 4.0 is based on the integration of production operating systems with information and communication technologies, where the Internet of Things (IoT) plays a key role. Modern production generates huge volumes of big data in the form of log files, signal streams and sensor data, which requires real-time systems for historical analysis and pattern recognition.

The use of a large number of sensors allows for continuous monitoring of the condition of machines, tools, and the quality of finished products. This allows for a drastic reduction in the level of unplanned stops.

Multimodal systems [2] combine data from different physical domains. For example, for electric motor diagnostics, the following are simultaneously analyzed:

- Vibration: allows you to detect rotor imbalance or bearing defects at an early stage.
- Temperature: indicates critical overheating of the windings or degradation of the lubricant.
- Acoustic emission: detects microcracks in metal long before visual or vibrational changes occur.

- Electrical current: Current signature analysis can detect electrical faults such as broken rotor bars.

The integration of these modalities allows the system to compensate for the limitations of each individual sensor. If the temperature sensor is inertial and slow to respond, the vibration sensor will detect the anomaly instantly, and the temperature indicator will confirm the severity of the problem, eliminating false alarms [3].

Multimodal systems are different in that they do not rely on a single source of information. They combine vibration, temperature, and acoustic emission data. This is critical because each individual physical quantity has its own limitations. For example, a temperature sensor may only detect overheating when damage has already begun, while vibration analysis will show microcracks much earlier. Combining these data allows for much higher accuracy compared to single-modal systems. Multimodal systems combine data from different sources, such as vibration, temperature, and acoustic emission, to compensate for the limitations of individual modalities and provide higher accuracy compared to single-modal systems.

The process of data fusion in machine learning is usually divided into early and late fusion. Early fusion occurs at the raw data or feature level. It allows deep learning models to see the relationships between different physical domains at the lowest level of abstraction. The effectiveness of multimodal data fusion systems in manufacturing relies critically on robust communication protocols and network architectures. To manage the vast data streams generated by the Industrial Internet of Things, the integration approach must prioritize low latency, scalability, and high reliability[4].

Communication Protocols:

- The MQTT [5] protocol is highly recommended for industrial environments, as it has proven to be up to six times faster than standard HTTP for transferring telemetry from sensors to central servers.
- Network Architecture: Implementing software-defined networking [6] enables the dynamic management of data flows, allowing the industrial system to adapt seamlessly to fluctuating execution conditions.
- Data Storage: Given that multimodal sensor data is inherently heterogeneous and unstructured, NoSQL [7] databases such as MongoDB or InfluxDB provide superior flexibility and scalability compared to traditional relational database management systems.
- Stream Processing: To analyze complex industrial processes, tools like Apache Storm, Apache Spark, and Apache Kafka [8] are utilized for processing large, real-time data streams, ensuring high reliability and parallel processing capabilities. Figure 2 demonstrates the key principle of identifying and distinguishing the modes, valleys, and clustering patterns is essential for designing robust deep learning architectures that can accurately model complex, multi-source data environments in smart manufacturing applications.

Each of the four sensor modalities: acoustic emission, thermal imaging, motor current, and operator maintenance logs — is processed by a dedicated deep learning submodel specifically designed for its signal type. Acoustic signals are converted into Mel-spectrograms and analyzed by a convolutional neural network to identify abnormal frequency patterns associated with mechanical defects. Thermal frames are processed by an image-based CNN that localizes hotspots and surface temperature anomalies[9]. Motor current signals undergo Wavelet decomposition followed by an LSTM network [10] to capture both time-frequency characteristics and their temporal evolution, enabling detection of gradual equipment degradation.

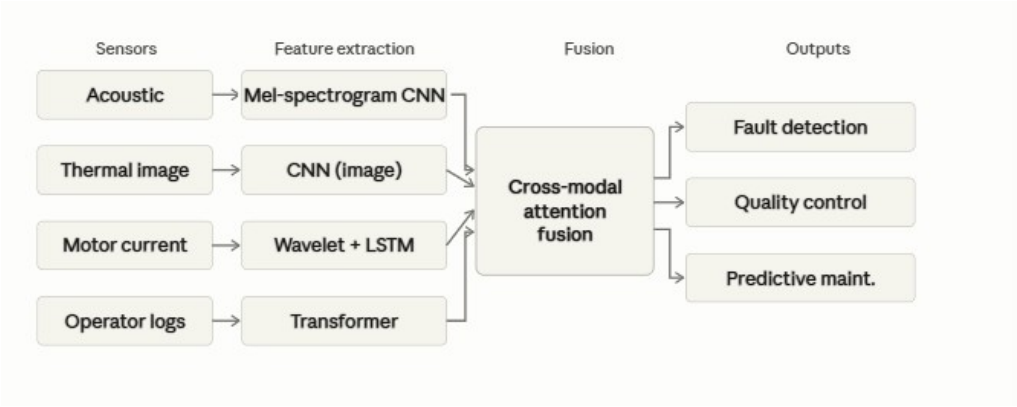

Figure 2: Multimodal Data Fusion for Predictive Modelling

Operator log texts are encoded by a Transformer-based language model [11] that extracts fault-relevant semantic context from maintenance records. Each submodel produces a compact feature vector of uniform dimensionality, representing the most informative characteristics of its respective modality. These four vectors are simultaneously passed to a cross-modal attention fusion module, which applies scaled dot-product attention to model dependencies between modalities for instance, allowing a thermal anomaly to reinforce a concurrent acoustic signal. An adaptive gating mechanism dynamically assigns weights [12] to each stream based on its current reliability, suppressing noisy or failed sensors and amplifying consistent ones, thereby ensuring robustness under real-world sensor degradation conditions.

The resulting fused representation is shared across three parallel output heads, each addressing a distinct manufacturing monitoring task: fault detection for identifying specific mechanical failures before breakdown, quality control for real-time surface defect classification, and predictive maintenance for estimating the remaining useful life of industrial equipment. This architecture enables a single unified model to simultaneously support multiple critical decision-making tasks in smart manufacturing environments.

3. Proposed approach

Our proposed approach utilizes a hybrid architecture equipped with cross-modal attention mechanisms to dynamically weight incoming data streams from various sources, such as video frames, acoustic signals, and text logs. The primary challenge is transforming this highly heterogeneous data into a shared latent space. By employing cross-modal attention, the model can reconcile features between different data sources for example, linking textual operator reports with thermal imaging. This attention mechanism [13] allows the system to focus on the most informative sensors at any given moment, significantly increasing resilience to transient sensor failures and high noise levels [14].

Data fusion in machine learning is typically classified as early (at the raw data or feature level) and late (at the algorithm output level). Early fusion allows models to learn relationships between domains at a low level, while late fusion [15] provides flexibility by using specialized classifiers for each modality.

Data collection and storage methods The effectiveness of data fusion systems critically depends on communication protocols and network architecture. The MQTT protocol has proven to be six times faster than HTTP for transferring data from sensors to servers, making it ideal for industrial environments. The use of software-defined networking allows for dynamic management of data flows and adaptation to changing execution conditions. Since sensor data is heterogeneous and unstructured, NoSQL databases such as MongoDB or InfluxDB prove to be more flexible and scalable compared to traditional relational systems. Tools such as Apache Storm, Apache Spark or Apache Kafka are often used to process large data streams in real time, providing high reliability and parallel processing.

Preprocessing and Model Architecture Sensor data [16] is often noisy, so filtering and feature extraction are critical steps in the analytical pipeline. Typical methods include vibration analysis using accelerometers, electrical current measurements, and thermography. Features are extracted in the time, frequency, and time-frequency domains, for example, through fast Fourier transforms or wavelet transforms.

Modern deep learning architectures use hybrid CNN-Transformer [17] models with attention mechanisms to dynamically weight the input streams. The attention mechanism allows the model to focus on the most informative sensors at a particular point in time, which increases resilience to transient failures of individual sensors. In addition, the use of cross-modal attention helps to reconcile features between different data sources, for example, by linking text reports with thermal images. Data obtained directly from the production site usually contains a lot of noise. Filtering and proper feature extraction are the most important steps in the analytical pipeline. Typical processing methods include vibration analysis using accelerometers and the motor current signature method.

Features are extracted in three main domains: time, frequency, and time-frequency. Time analysis gives an overall view of the signal energy, while frequency analysis through fast Fourier transform allows you to identify specific frequencies corresponding to certain defects. Wavelet transform adds the ability to see how the signal spectrum changes over time, which is indispensable for non-stationary processes.

3.1. Data Fusion Architectures

Let $f_m \in \mathbb{R}^d_m$ represent the extracted feature vector for modality $m \in \{1, 2, ..., M\}$.

Models learn relationships at a low level by concatenating raw or preprocessed features before classification [7]:

$$F_{early} = [f_1 \oplus f_2 \oplus \cdots \oplus f_M] \quad (1)$$

$$y_{pred} = \text{Classifier}(F_{early}; \theta_c) \quad (2)$$

Uses specialized classifiers C_m for each modality and aggregates their output probabilities or decisions [**3, 4**]:

$$y_{late} = \sum_{m=1}^{M} w_m C_m(f_m) \quad (3)$$

where w_m represents the static or learned weight for modality m.

To address sensor failures, high noise, and “modal bias”, the system maps heterogeneous data to a shared latent space and dynamically weights the streams [8].

Each feature vector f_m is projected into a common dimension d [5, 6]:

$$h_m = \sigma(W_m f_m + b_m) \quad (4)$$

where W_m is the learned weight matrix, b_m is the bias, and σ is an activation function (e.g., ReLU).

To reconcile features between different modalities (e.g., linking textual operator reports (modality A) with thermal imaging (modality B)), Scaled Dot-Product Attention is utilized. Let $Q_a = h_a W_k$ (Query from Text), and $K_n = h_n W_k$, $V_n = h_n WV$(Key and Value from Thermal).

$$\text{Attention}(A \rightarrow B) = \text{Softmax}\left(\frac{Q_A K_B^T}{\sqrt{d_k}}\right) V_B \quad (5)$$

The model dynamically assesses the informativeness of each sensor at time t to maintain resilience against transient failures [18]. An attention score (gating value) α_m is computed for each modality:

$$\alpha_m = \text{Softmax}(W_{gate} h_m + b_{gate}) = \frac{\exp(W_{gate} h_m + b_{gate})}{\sum_{j=1}^{M} \exp(W_{gate} h_j + b_{gate})} \quad (6)$$

The final fused representation is the dynamically weighted sum of the latent vectors:

$$H_{fusion} = \sum_{m=1}^{M} \alpha_m h_m \quad (7)$$

3.2. Explainable AI Approach

To ensure transparency for critical processes, the decisions of the deep learning model $f(x)$ are explained using additive feature attribution methods (like SHAP or LIME) [19].

The complex model $f(x)$ is locally approximated by an interpretable linear model $g(z')$:

$$g(z') = \varphi_0 + \sum_{i=1}^{P} \varphi_i z_i' \quad (8)$$

where $z'_i \in \{0, 1\}$ indicates the presence of a specific multimodal feature, and φ_i is the calculated Shapley value representing the influence (weight) of parameter i on the failure prediction.

To safely operate during data distribution shifts, the model calculates a confidence interval or variance score $\sigma^2 epistemic$ associated with the fused features, ensuring the cascaded detection architecture can flag when the model is over-relying on a weaker, noisy modality [20].

In the field of 3D printing, multimodal integration of acoustic, vibration and thermal signals allows the detection of defects such as nozzle clogging or filament runout with an accuracy of over 90%. Experiments show that acoustic signals are sensitive to extrusion problems, while accelerometers are better at detecting mechanical motion anomalies. In laser metal deposition processes, the combination of acoustic emission signals with thermal images of the melt pool allows the prediction of porosity at the submillimeter level of surface features [21].
AE sensors capture elastic waves from laser-material interaction, which makes it possible to identify metal spatter and areas of insufficient melting. Another important direction is the creation of digital twins for gearbox diagnostics, where data from real sensors is combined with virtual simulations to reduce the information gap [22].

To implement AI in critical processes, it is necessary to ensure transparency of the decisions made through xAI methods. Tools such as LIME or SHAP help engineers understand which parameters influenced the failure prediction. Analysis of gate weights in multimodal models reveals “modal bias”, where the system may over-rely on the weaker modality, requiring cascaded detection architectures. New approaches such as hyper dimensional computing allow for efficient assessment of epistemic uncertainty at the feature level with minimal computational overhead. This ensures robust performance of models even in challenging environments with high noise levels or with data distribution shifts.
The main challenge remains the synchronization of data coming in at different sampling rates. Combining a 20 kHz vibration signal and temperature readings updated once per second requires

complex mathematical alignment algorithms. The future of the industry lies in the realm of edge computing, where part of the analysis takes place directly on the controller at the machine. This will allow for instant response to critical situations without wasting time transferring data to the cloud. Thus, the integration of multimodal data not only improves diagnostics, but also lays the foundation for fully autonomous factories of the future.

Conclusions

The transition toward Industry 4.0 relies fundamentally on the seamless integration of operational technologies and advanced information systems. As demonstrated in this study, relying on single-modal sensor data is insufficient for the complex, high-stakes environments of modern manufacturing. The integration of multimodal data encompassing vibration, temperature, acoustic emission, textual logs, and electrical current provides a robust, fault-tolerant framework capable of overcoming individual sensor limitations, data distribution shifts, and environmental noise.

The realization of autonomous condition monitoring and diagnostics depends on several critical pillars:

1. Advanced Fusion Architectures: The implementation of hybrid deep learning models utilizing cross-modal attention mechanisms allows for the dynamic weighting of heterogeneous data streams. By mapping diverse data into a shared latent space, the system can adaptively focus on the most reliable sensors, ensuring continuous predictive diagnostics even during transient sensor failures.
2. Explainable AI (xAI): For intelligent systems to be adopted in critical manufacturing processes, operational transparency is non-negotiable. Integrating additive feature attribution methods like SHAP and LIME provides necessary interpretability, mitigating the risks of modal bias and ensuring that human operators can trust and verify predictive maintenance alerts.
3. Robust IoT Infrastructure: The sheer volume and velocity of industrial data necessitate specialized network architectures. Utilizing protocols like MQTT, NoSQL databases, and software-defined networking ensures the low latency and scalability required for real-time stream processing.

Moving forward, addressing the complex synchronization of multimodal data with varying sampling rates remains a primary challenge. However, as the industry shifts toward edge computing processing data directly on machine controllers rather than relying entirely on cloud infrastructure these latency and synchronization barriers will diminish. Ultimately, the deep integration of multimodal data fusion not only enhances immediate diagnostic accuracy but serves as the foundational architecture for the fully autonomous, self-healing factories of the future.

Acknowledgements

This study was funded by the National Research Foundation of Ukraine in the framework of the research project 2025.07/0017 on the topic “Methods of analysis and optimization of multimodal data for deep learning models in the military sphere”.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini Pro Instant and Grammarly in order to: text polishing, grammar and spelling check. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] Ghobakhloo, M. (2020). Industry 4.0, digitization, and opportunities for sustainability. Journal of cleaner production, 252, 119869.

[2] Tsanousa, A., Bektsis, E., Kyriakopoulos, C., González, A. G., Leturiondo, U., Gialampoukidis, I., & Kompatsiaris, I. (2022). A review of multisensor data fusion solutions in smart manufacturing: Systems and trends. Sensors, 22(5), 1734.

[3] Rana, S. (2025). AI-driven fault detection and predictive maintenance in electrical power systems: A systematic review of data-driven approaches, digital twins, and self-healing grids. American Journal of Advanced Technology and Engineering Solutions, 1(01), 258-289.

[4] Kopetz, H., & Steiner, W. (2022). Internet of things. In Real-time systems: design principles for distributed embedded applications (pp. 325-341). Cham: Springer International Publishing.

[5] Quincozes, S., Emilio, T., & Kazienko, J. (2019). MQTT protocol: fundamentals, tools and future directions. IEEE Latin America Transactions, 17(09), 1439-1448.

[6] Cox, J. H., Chung, J., Donovan, S., Ivey, J., Clark, R. J., Riley, G., & Owen, H. L. (2017). Advancing software-defined networks: A survey. Ieee Access, 5, 25487-25526.

[7] Lakshminarayana, S., Praseed, A., & Thilagam, P. S. (2024). Securing the IoT application layer from an MQTT protocol perspective: Challenges and research prospects. IEEE Communications Surveys & Tutorials, 26(4), 2510-2546.

[8] Vyas, S., Tyagi, R. K., Jain, C., & Sahu, S. (2021, July). Literature review: A comparative study of real time streaming technologies and apache kafka. In 2021 Fourth International Conference on Computational Intelligence and Communication Technologies (CCICT) (pp. 146-153). IEEE.

[9] Bulgin, C. E., Merchant, C. J., & Ferreira, D. (2020). Tendencies, variability and persistence of sea surface temperature anomalies. Scientific reports, 10(1), 7986.

[10] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica d: Nonlinear phenomena, 404, 132306.

[11] Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2021). Ammus: A survey of transformer-based pretrained models in natural language processing. arXiv preprint arXiv:2108.05542.

[12] Zhang, G., Zhang, S., Bachras, M., Zhang, Y., & Jacobsen, H. A. (2025). Cabinet: Dynamically Weighted Consensus Made Fast. arXiv preprint arXiv:2503.08914.

[13] Marrah, SA, et al. Deep Learning-Based Adaptive Sensor Fusion for Real-Time Control and Fault-Tolerant Automation in IoT Systems. Preprints.org (2026).

[14] Picaut, J., Can, A., Fortin, N., Ardouin, J., & Lagrange, M. (2020). Low-cost sensors for urban noise monitoring networks—A literature review. Sensors, 20(8), 2256.

[15] Gadzicki, K., Khamsehashari, R., & Zetzsche, C. (2020, July). Early vs late fusion in multimodal convolutional neural networks. In 2020 IEEE 23rd international conference on information fusion (FUSION) (pp. 1-6). IEEE.

[16] Karthikeyan, A., et al. In-situ surface porosity prediction in hybrid-directed energy deposition process using explainable multimodal sensor fusion. ArXiv (2024).

[17] Khan, A., Rauf, Z., Sohail, A., Khan, A. R., Asif, H., Asif, A., & Farooq, U. (2023). A survey of the vision transformers and their CNN-transformer based variants. Artificial Intelligence Review, 56(Suppl 3), 2917-2970.

[18] Giannakopoulos, I., Konstantinou, I., Tsoumakos, D., & Koziris, N. (2018). Cloud application deployment with transient failure recovery. Journal of Cloud Computing, 7(1), 11.

[19] Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 7(1), 2400304.

[20] Liao, C., Lei, K., Zheng, X., Moon, J., Wang, Z., Wang, Y., ... & Hu, X. (2025). Benchmarking multi-modal semantic segmentation under sensor failures: Missing and noisy modality robustness. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 1576-1586).

[21] Malekzadeh, M., Clegg, R., Cavallaro, A., & Haddadi, H. (2021). Dana: Dimension-adaptive neural architecture for multivariate sensor data. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5(3), 1-27.

[22] Tammisetti, A.K.; Nalamalapu, K.S.; Nagella, S.; Shaik, K.; Shaik, K.A. Deep Residual Learning based Attendance Monitoring System. In Proceedings of the 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 25-26 March 2022; pp. 1089-1093.