

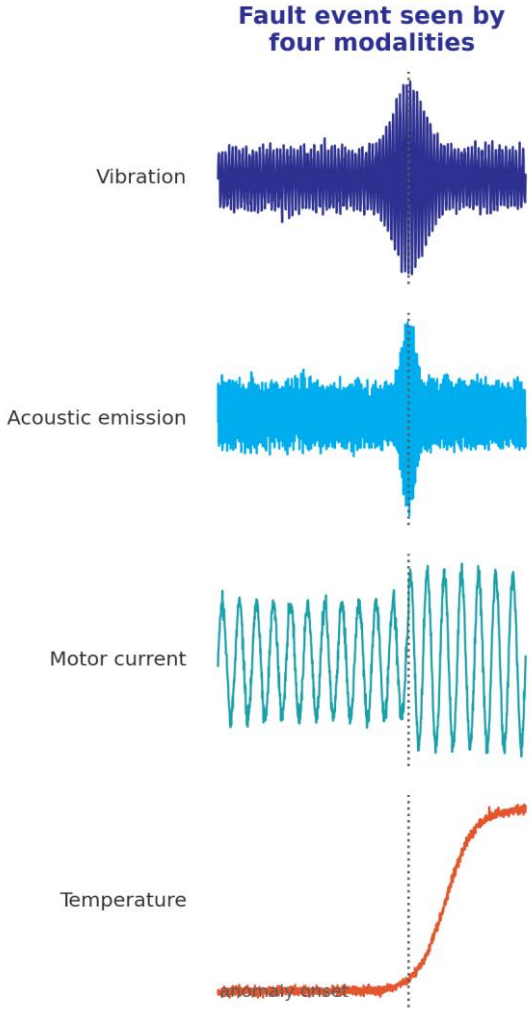
Multimodal Data Fusion Approaches in Intelligent Systems

Oleh Basystiuk

HORIZON TMA MSCA Doctoral Networks • HORIZON-MSCA-2023-DN-01 • Gliwice, Poland, 20-23 July 2026

Why multimodal?

- Industry 4.0 fuses operational technology with information systems — the Industrial IoT is its core enabler.
- A modern shop floor streams big data continuously: log files, signal streams and telemetry.
- Every single sensor is limited: temperature reacts late, vibration reacts instantly, acoustics react earliest.
- Multimodal fusion compensates for the blind spots of each modality and suppresses false equipment responses.
- Objective: fault-tolerant and explainable deep learning for autonomous condition monitoring.



Structure of multimodal data

- Mode 1 — the highest peak: the dominant trend, the most frequent cluster of values.
- Mode 2 — a secondary peak: an alternative trend or a subpopulation in the data.
- Valleys — low-frequency regions between peaks; natural boundaries between modal clusters.
- Clustering patterns — the broadest level: the overall distributional structure of the dataset.



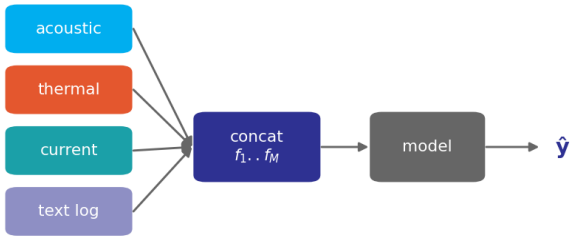
Fusion strategies

Early • Late • Hybrid adaptive weighting

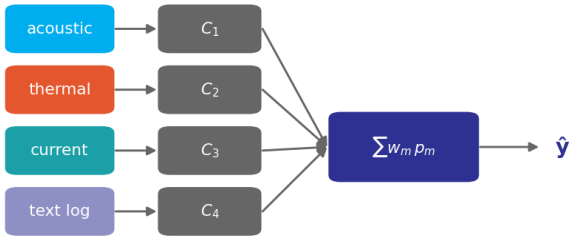
Early fusion concatenates raw or preprocessed feature vectors f_m before classification, so the model learns relationships between physical domains at the lowest level of abstraction. Late fusion keeps a specialised classifier C_m per modality and aggregates their outputs with static or learned weights w_m — more flexible, but blind to low-level interactions.

The proposed hybrid architecture projects every f_m into a shared latent space $h_m = \sigma(W_m f_m + b_m)$ of common dimension d and weights the streams dynamically, combining the sensitivity of early fusion with the fault tolerance of late fusion.

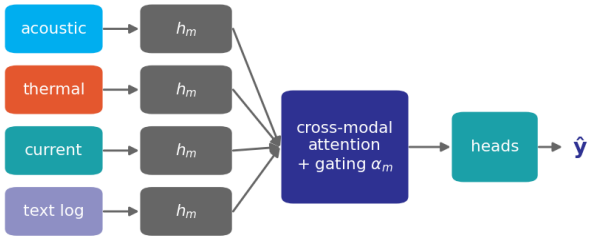
Early fusion — feature level



Late fusion — decision level

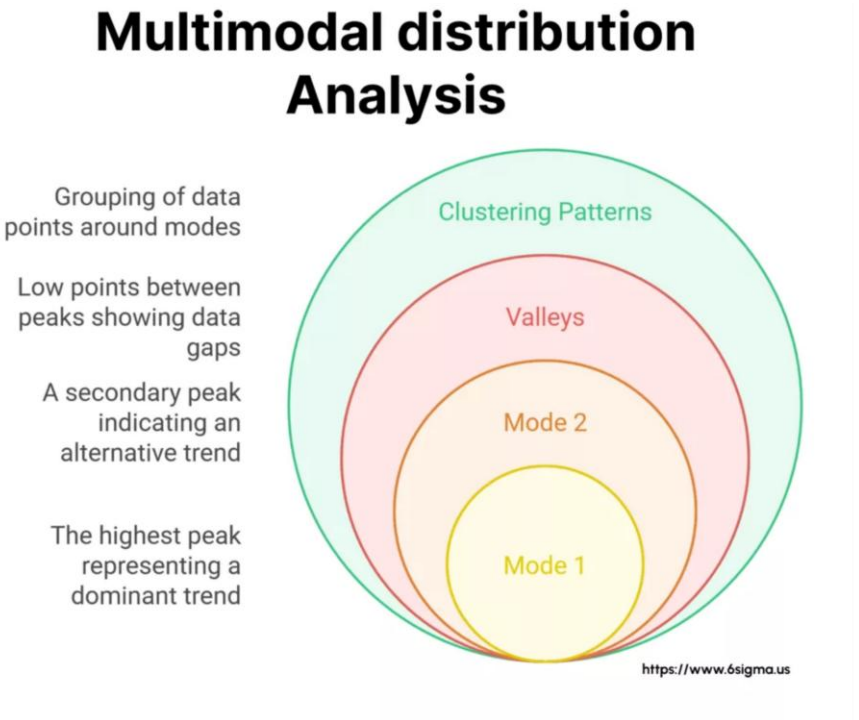


Hybrid — adaptive weighting (proposed)



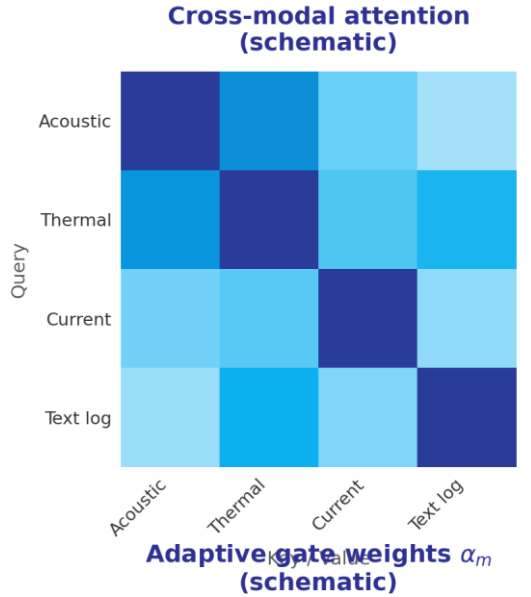
Proposed architecture

- Acoustic emission → Mel-spectrogram + CNN
- Thermal frames → image CNN localising hotspots
- Motor current → wavelet + LSTM
- Operator logs → Transformer language model
- Each submodel emits a compact feature vector of uniform dimensionality, fused by cross-modal attention.



Cross-modal attention

- Heterogeneous features are mapped into one shared latent space of dimension d.
- Scaled dot-product attention reconciles modalities — e.g. a textual operator report (Query) with a thermal frame (Key, Value).
- An adaptive gating mechanism computes a score α_m for each stream from its current reliability.
- Noisy or failed sensors are suppressed, consistent ones amplified — the fused vector is their weighted sum.
- Result: resilience to transient sensor failures, high noise and modal bias.





Three monitoring tasks

The fused representation is shared across three parallel output heads, so a single unified model supports several critical manufacturing decisions at once.

Fault detection



Identifies specific mechanical failures before breakdown.

Quality control



Real-time classification of surface defects on the line.

Predictive maintenance



Estimates the remaining useful life of the equipment.

Data infrastructure

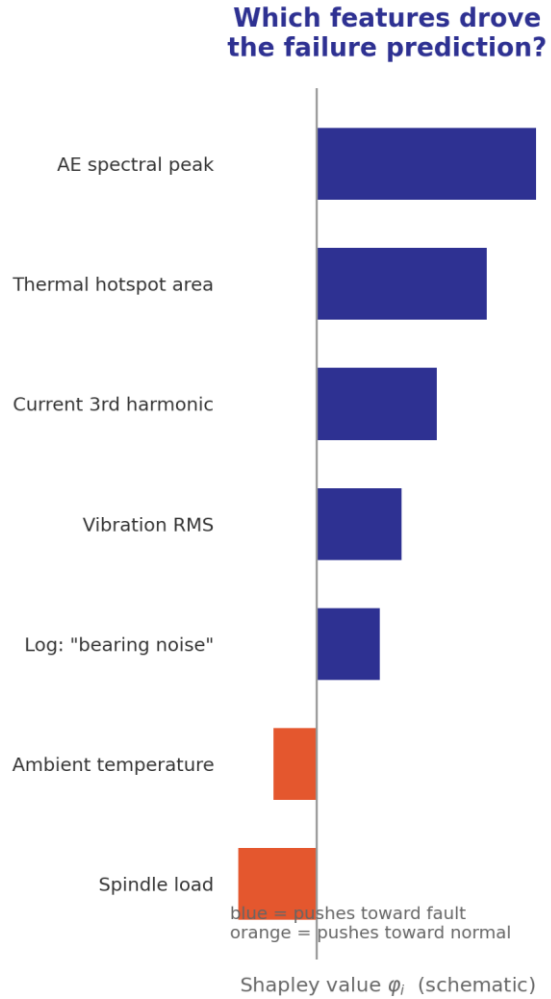
Protocols, network, storage

MQTT moves telemetry from sensors to servers up to six times faster than HTTP, which makes it the default choice for industrial environments. Software-defined networking manages data flows dynamically and lets the system adapt to fluctuating execution conditions.

Sensor data is heterogeneous and unstructured, so NoSQL stores such as MongoDB or InfluxDB scale better than relational systems. Apache Kafka, Spark and Storm process large real-time streams in parallel with high reliability.

Preprocessing and features

Data from the production site is noisy, so filtering and feature extraction are the most important steps of the analytical pipeline: vibration analysis with accelerometers, motor current signature analysis, thermography. Features are extracted in three domains — time (overall signal energy), frequency (FFT isolates defect-specific components) and time–frequency (wavelets show how the spectrum evolves, indispensable for non-stationary processes).



Explainable AI

Transparency for critical processes

The decisions of the deep model f(x) are explained with additive feature attribution: f(x) is locally approximated by an interpretable linear model g(z′), where the Shapley value ϕ_i measures the influence of each multimodal feature on the failure prediction. SHAP and LIME let engineers see which parameters drove the alarm.

Analysis of the gate weights exposes modal bias — over-reliance on a weaker modality — which motivates cascaded detection architectures. An epistemic variance score flags data distribution shifts, and hyperdimensional computing estimates uncertainty at feature level with minimal computational overhead.

Industrial applications

Multimodal integration has already been validated across several manufacturing domains, from desktop extrusion to metal additive processes.

3D printing



Acoustic, vibration and thermal signals detect nozzle clogging or filament runout with over 90% accuracy.

Laser metal deposition



Acoustic emission with melt-pool thermal images predicts porosity at submillimetre level.

Digital twins



Real sensor data merged with virtual simulation closes the information gap in gearbox diagnostics.

Conclusions

- A late and adaptive weighted fusion architecture on transformers and DNNs raises diagnostic accuracy and fault tolerance.
- Cross-modal attention with adaptive gating keeps the system stable under real-world sensor degradation.
- SHAP / LIME attribution and epistemic uncertainty estimates make the decisions auditable in critical processes.
- Open challenge: synchronising a 20 kHz vibration signal with temperature readings updated once per second.
- Next step — edge computing on the machine controller: instant reaction without a cloud round-trip.



Thank you for your attention!

Oleh Basystiuk — oleh.a.basystiuk@lpnu.ua
http://lpnu.ua