International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(10s) 2026 pp. 405-420
https://doi.org/10.70917/ijcisim-2026-3594
Article

Multimodal Deep Learning for Intelligent Condition Monitoring Using Vibration, Acoustic and Thermal Data

B. C. Anilkumar[1], Anver S. R[2], Rahul C[3], M. Sarith Divakar[4], Arathi T[5]

[1]Department of Mechanical Engineering, LBS College of Engineering Kasaragod.
Email: anilkbc@lbscek.ac.in
[2]Department of Computer Science & Engineering, LBS College of Engineering Kasaragod.
Email: anversr@lbscek.ac.in
[3]Department of Computer Science & Engineering, LBS College of Engineering Kasaragod.
Email: rahul@lbscek.ac.in
[4]Department of Computer Science and Engineering, LBS College of Engineering, Kasaragod.
Email: sarith@lbscek.ac.in
[5]Department of Electronics & Communication Engineering, LBS College of Engineering Kasaragod.
Email: arathit@lbscek.ac.in

Abstract: Industrial condition monitoring increasingly relies on heterogeneous sensors because no single modality completely captures the mechanical, acoustic, and thermal signatures of developing faults. Vibration signals are highly sensitive to impacts, resonance, imbalance, and misalignment; acoustic data provide non-contact evidence of friction, leakage, looseness, and impulsive events; and infrared thermal images reveal heat accumulation, abnormal resistance, lubrication loss, and overload. This study presents a multimodal deep-learning framework that integrates these three sources through modality-specific convolutional encoders and a sample-wise reliability gate. Each vibration and acoustic window is processed by a compact one-dimensional convolutional neural network, while thermal maps are processed by a two-dimensional convolutional branch. The resulting embeddings are fused through softmax-normalized modality weights and a shared representation before four-class diagnosis. A reproducible synthetic benchmark containing 1,200 observations was generated for healthy operation, bearing fault, misalignment, and overheating. The data were divided into 840 training, 180 validation, and 180 test observations; the principal test set included an additional noise shift to approximate sensor degradation and field variability. The proposed attentive fusion network achieved 99.44% accuracy and a macro F1-score of 0.9944, outperforming vibration-only (76.11%), acoustic-only (88.89%), thermal-only (54.44%), and late probability fusion (82.22%) baselines. Noise and missing-modality tests showed that performance declined as corruption increased and that vibration and thermal information were especially important in the present synthetic design. The results demonstrate the methodological value of reliability-aware intermediate fusion, while also emphasizing that synthetic evidence is not a substitute for industrial validation. The paper provides an implementation-oriented architecture, evaluation protocol, deployment blueprint, and research agenda for robust multimodal predictive maintenance.

Keywords: condition monitoring; multimodal deep learning; vibration analysis; acoustic monitoring; infrared thermography; sensor fusion; predictive maintenance; fault diagnosis

1. Introduction

Condition-based maintenance seeks to schedule inspection and intervention according to the observed health of equipment rather than fixed calendar intervals. A typical condition-monitoring chain includes sensing, signal processing, diagnosis, prognosis, and maintenance decision support. Reviews of the field have consistently shown that the value of a CBM program depends not only on diagnostic algorithms but also on data quality, contextual information, uncertainty management, and the practical link between model outputs and maintenance actions (Jardine et al., 2006; Lei et al., 2018).

Rotating machines, pumps, fans, motors, gearboxes, and process equipment generate several physical manifestations of degradation. A bearing defect may produce impulsive vibration, high-frequency acoustic bursts, and a localized temperature rise. Misalignment may create harmonics and directional vibration while simultaneously changing sound radiation and heat distribution. Electrical resistance, blocked ventilation, overload, and lubrication failure often emerge thermally before catastrophic breakdown. These signatures are related but not identical; therefore, relying on a single modality can produce blind spots when the signal-to-noise ratio changes or a sensor is poorly positioned.

Deep learning has reduced dependence on manually designed features by learning representations directly from raw or minimally processed signals. One-dimensional CNNs have been effective for vibration and motor-current diagnosis (Ince et al., 2016; Kiranyaz et al., 2021), two-dimensional CNNs can extract spatial temperature patterns from thermal images, and learned acoustic representations can support anomaly detection in industrial machines (Purohit et al., 2019; Koizumi et al., 2019). However, multimodal learning introduces additional questions: how should heterogeneous sampling rates be synchronized, where should information be fused, how should unreliable sensors be down-weighted, and how should the system behave when one modality is missing?

This paper addresses these questions through a reliability-aware intermediate-fusion architecture for vibration, acoustic, and thermal data. The focus is methodological rather than application-specific. A synthetic benchmark is used so that the complete data-generation and evaluation process can be described without claiming access to industrial measurements. The benchmark intentionally combines complementary condition signatures and includes a noise-shifted test set. This design supports a controlled comparison of unimodal, pairwise, late-fusion, and attentive-fusion alternatives.

The principal contributions are: (1) an end-to-end multimodal architecture with specialized 1-D and 2-D encoders; (2) a softmax reliability gate that learns sample-wise modality weights; (3) a reproducible four-condition synthetic benchmark and transparent experimental protocol; (4) ablation studies for sensor combinations, noise, and missing modalities; and (5) an industrial deployment blueprint covering synchronization, edge inference, uncertainty, cybersecurity, and maintenance integration.

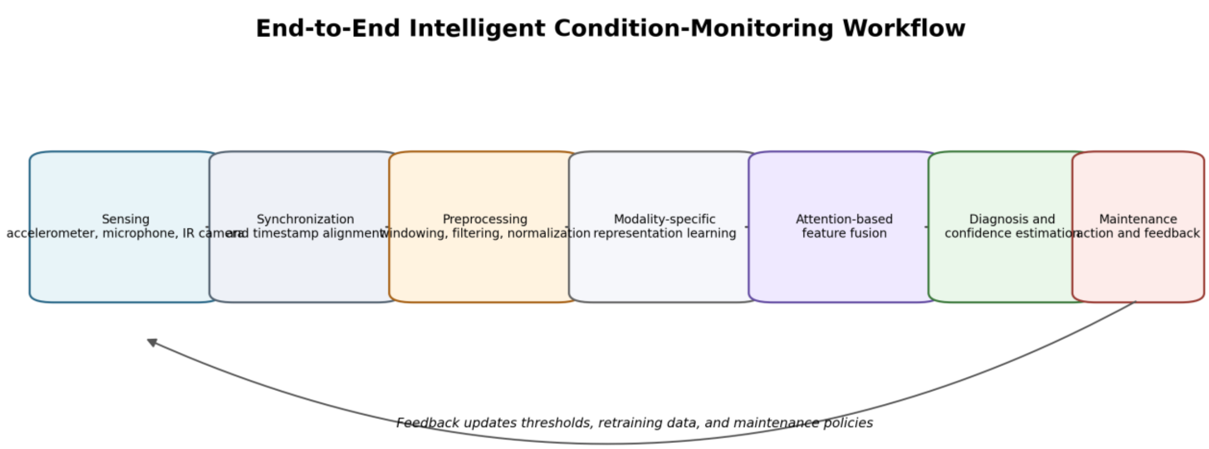

Figure 1. End-to-end workflow for multimodal intelligent condition monitoring.

2. Literature Review and Theoretical Background

2.1 Condition monitoring and prognostics

The modern PHM literature distinguishes diagnosis, which estimates the current fault state, from prognosis, which estimates future degradation and remaining useful life. Jardine et al. (2006) organized CBM around data acquisition, data processing, and decision-making, while Lei et al. (2018) reviewed the full prognostic chain from sensing to health indicators and RUL prediction. These frameworks imply that a high classification score is insufficient if the model is not calibrated, traceable, and linked to maintenance policy.

Deep-learning reviews report strong performance across bearing, gearbox, motor, and tool-health applications but also identify recurring weaknesses: limited cross-machine generalization, inconsistent benchmark protocols, class imbalance, domain shift, and insufficient reporting of computational cost (Khan & Yairi, 2018; Zhao et al., 2019). Recent work on deep generative models further highlights opportunities for data augmentation, domain adaptation, uncertainty modeling, and multimodal fusion, while warning that generated data may conceal physical inconsistencies or produce overconfident models (Yang et al., 2026).

2.2 Vibration-based monitoring

Vibration analysis remains a central technique because mechanical faults alter force transmission, resonance, modulation, and impulsiveness. Classical bearing diagnosis uses characteristic frequencies, envelope analysis, cyclostationary analysis, and spectral kurtosis to isolate fault-related transients (Antoni, 2006; Randall & Antoni, 2011). The CWRU benchmark study by Smith and Randall (2015) also demonstrated that widely used public records vary substantially in diagnostic difficulty and should not be treated as a uniform gold standard.

CNNs learn local filters that can detect impulsive and harmonic patterns directly from waveforms or time-frequency images. Ince et al. (2016) demonstrated real-time motor fault detection using compact 1-D CNNs, Janssens et al. (2016) used CNN-based representation learning for rotating machinery, and Wen et al. (2018) converted signals to two-dimensional forms for data-driven diagnosis. One-dimensional architectures are attractive for edge deployment because they can avoid the cost of spectrogram generation and use fewer parameters than large image networks (Kiranyaz et al., 2021).

2.3 Acoustic monitoring

Acoustic sensing is non-contact and can capture leakage, rubbing, impacts, cavitation, looseness, and aerodynamic disturbances. It is especially useful when accelerometers cannot be mounted or when a single microphone can observe several assets. Acoustic measurements are nevertheless vulnerable to reflections, competing machines, microphone directionality, and changing room acoustics.

The MIMII dataset introduced realistic normal and anomalous sounds for valves, pumps, fans, and slide rails, while ToyADMOS provided controlled miniature-machine operating sounds for anomaly-detection research (Purohit et al., 2019; Koizumi et al., 2019). General audio CNN research has shown that convolutional networks can learn meaningful spectro-temporal patterns from large sound corpora (Piczak, 2015; Hershey et al., 2017). In industrial systems, robust acoustic monitoring typically requires augmentation for background noise, reverberation, gain changes, and operating-speed variation.

2.4 Infrared thermal monitoring

Infrared thermography provides non-contact spatial temperature measurement. The review by Bagavathiappan et al. (2013) documented applications across mechanical, electrical, civil, and process systems and emphasized the importance of emissivity, viewing angle, reflections, distance, ambient temperature, and acquisition procedures. Thermal images can expose abnormal resistance, friction, poor lubrication, overload, blocked cooling, and asymmetry that may not be obvious in vibration data.

Huda and Taib (2013) demonstrated the value of thermography for predictive and preventive maintenance of electrical defects, and Glowacz and Glowacz (2017) used thermal-image features for induction-motor diagnosis. Thermal monitoring is often slower than vibration or acoustic sensing because heat generation and diffusion have longer time constants. This difference can be advantageous: thermal data provide a complementary view of accumulated energy rather than instantaneous dynamics.

2.5 Multimodal learning and fusion

Multisensor fusion combines observations to improve completeness, robustness, and uncertainty handling. Hall and Llinas (1997) and Khaleghi et al. (2013) described fusion as a broad family of methods operating at data, feature, or decision levels. Multimodal machine learning extends this perspective by addressing representation, alignment, fusion, translation, and co-learning (Baltrusaitis et al., 2019). Deep multimodal surveys distinguish early fusion, intermediate fusion, and late fusion, with the central trade-off being the degree of cross-modal interaction versus modularity and robustness (Ramachandram & Taylor, 2017).

Early fusion concatenates aligned raw measurements or low-level features. It is simple but sensitive to scale and synchronization. Late fusion combines independently generated probabilities or decisions and therefore supports modular deployment, but it may miss interactions that become visible only in a joint representation. Intermediate fusion learns modality-specific embeddings and then models their relationships. Gated multimodal units, tensor fusion, and modality dropout have been proposed to manage variable modality reliability and cross-modal structure (Arevalo et al., 2017; Neverova et al., 2016; Zadeh et al., 2017).

The present study selects gated intermediate fusion because vibration, acoustic, and thermal data differ in dimensionality, sampling rate, physical meaning, and noise processes. A shared encoder would force incompatible inputs into one representation too early, whereas pure late fusion would not directly model feature-level complementarity.

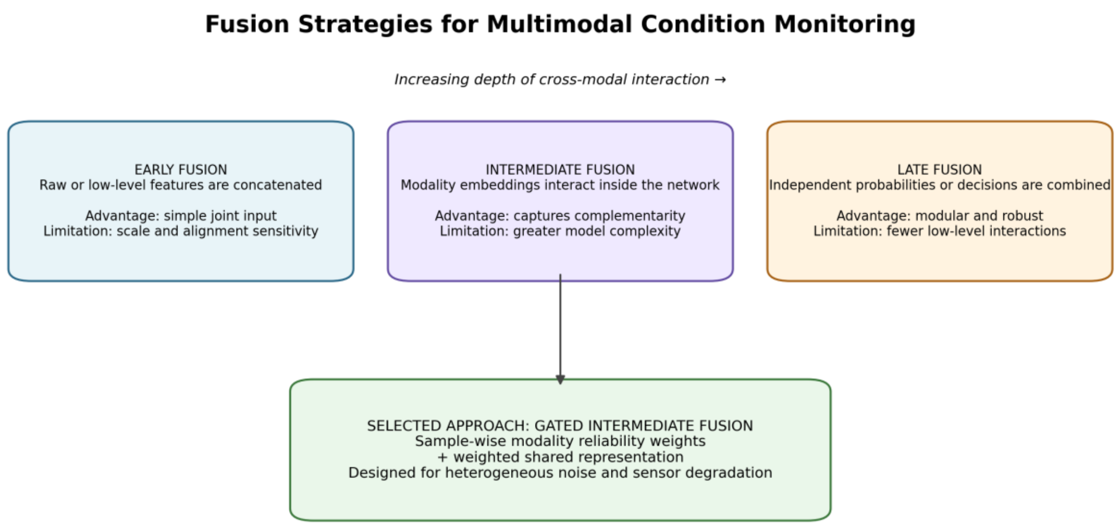

Figure 2. Comparison of early, intermediate, and late fusion strategies.

2.6 Representative literature and research gap

Study	Focus	Modality	Contribution	Gap relevant to this paper
Jardine et al. (2006)	CBM review	Multiple	Links sensing, diagnosis, prognosis, and decisions	Predates modern deep multimodal learning
Randall & Antoni (2011)	Bearing diagnostics	Vibration	Mechanistic tutorial on impulsive bearing signatures	Single modality
Ince et al. (2016)	Motor fault diagnosis	Vibration/current	Compact real-time 1-D CNN	No thermal or acoustic fusion
Janssens et al. (2016)	Rotating machinery	Vibration	CNN reduces manual feature engineering	Single modality

Purohit et al. (2019)	Industrial anomaly detection	Acoustic	Realistic machine-sound dataset	No thermal/vibration pairing
Koizumi et al. (2019)	Acoustic anomaly detection	Acoustic	Large controlled dataset	Miniature machines
Bagavathiappan et al. (2013)	Condition monitoring review	Thermal	Comprehensive IRT principles and applications	Limited deep-learning focus
Glowacz & Glowacz (2017)	Motor diagnosis	Thermal	Fault recognition from motor thermal images	Feature-based and single modality
Baltrusaitis et al. (2019)	Multimodal ML survey	General	Taxonomy of representation, alignment, and fusion	Not specific to industrial CM
Neverova et al. (2016)	Adaptive multimodal learning	Multiple	Modality dropout and adaptive fusion	Different application domain
Zhao et al. (2019)	Machine health review	Mostly vibration	Organizes DL methods for health monitoring	Limited tri-modal evaluation
Yang et al. (2026)	Generative PHM review	Multiple	Future directions for generation, fusion, and adaptation	Evidence still fragmented

Table 1. Representative literature and the gap addressed by the proposed tri-modal framework.

The literature contains strong unimodal methods and general multimodal theory, but comparatively few studies provide an integrated and transparent evaluation of vibration, acoustic, and thermal information using a reliability-aware fusion mechanism. Public datasets are usually modality-specific and rarely synchronized across all three sensors. The research gap is therefore both architectural and experimental: a practical framework must preserve modality-specific structure, model cross-modal complementarity, and explicitly test noise and sensor loss.

3. Research Questions, Hypotheses and Contributions

RQ1: Does intermediate attentive fusion improve condition classification relative to vibration-only, acoustic-only, thermal-only, pairwise, and late-fusion baselines?

RQ2: How does the proposed model behave when sensor noise increases?

RQ3: Which modalities are most influential in the controlled synthetic benchmark, and what happens when one modality is unavailable?

RQ4: What implementation controls are required to translate a proof-of-concept model into an industrial condition-monitoring system?

The principal hypothesis is that a trained reliability gate can exploit complementary information more effectively than fixed probability averaging. A secondary hypothesis is that performance will decline gracefully under moderate corruption but will deteriorate when a dominant modality is removed or when noise exceeds the training distribution.

4. Materials and Methods

4.1 Study design and reproducibility

The study is a controlled proof-of-concept experiment. All observations were generated algorithmically using a fixed random seed (42), and all model comparisons used the same training, validation, and test partitions. The objective is not to claim industrial performance but to verify the behavior of a tri-modal deep architecture under known synthetic conditions. The full benchmark contained 1,200 observations, with 300 observations in each of four classes: healthy, bearing fault, misalignment, and overheating. Stratified partitioning produced 840 training observations, 180 validation observations, and 180 test observations.

4.2 Synthetic multimodal data generation

Each observation contained a 512-sample vibration waveform, a 512-sample acoustic waveform, and a 24 x 24 thermal map. The waveform generator included a fundamental component, harmonics, amplitude variability, phase variability, load variability, speed variability, and Gaussian noise. The thermal generator included a baseline temperature field, a weak spatial gradient, pixel noise, and class-dependent hotspot patterns. The class mechanisms were designed as follows:

Healthy: Fundamental and weak harmonic content, normal temperature field, and occasional benign perturbation.

Bearing fault: Repeated impulsive excitation with decaying ringing in vibration and acoustic channels, plus a localized thermal hotspot.

Misalignment: Increased second and third harmonics in vibration and acoustic channels, plus asymmetric thermal hotspots.

Overheating: Subtle waveform changes with increased broadband noise and a broad temperature rise with diffuse heating.

Waveforms were standardized per observation to focus the temporal encoders on shape and spectral composition rather than absolute gain. Thermal maps were standardized globally so that relative temperature structure and spatial gradients remained available. The principal test loader added Gaussian corruption with a standard deviation of 0.30 to the waveform channels and a proportionally smaller disturbance to thermal maps. This produces a deliberate train-test shift and prevents a trivially clean evaluation.

Parameter	Value	Purpose/interpretation
Classes	4	Healthy, bearing fault, misalignment, overheating
Observations	1,200	300 per class
Waveform length	512 samples	Vibration and acoustic
Thermal resolution	24 x 24 pixels	Single-channel standardized map
Training/validation/test	840 / 180 / 180	Stratified by class
Main test noise	0.30	Gaussian corruption added at evaluation
Random seed	42	Used for generation, partitioning, and training

Table 2. Synthetic benchmark configuration.

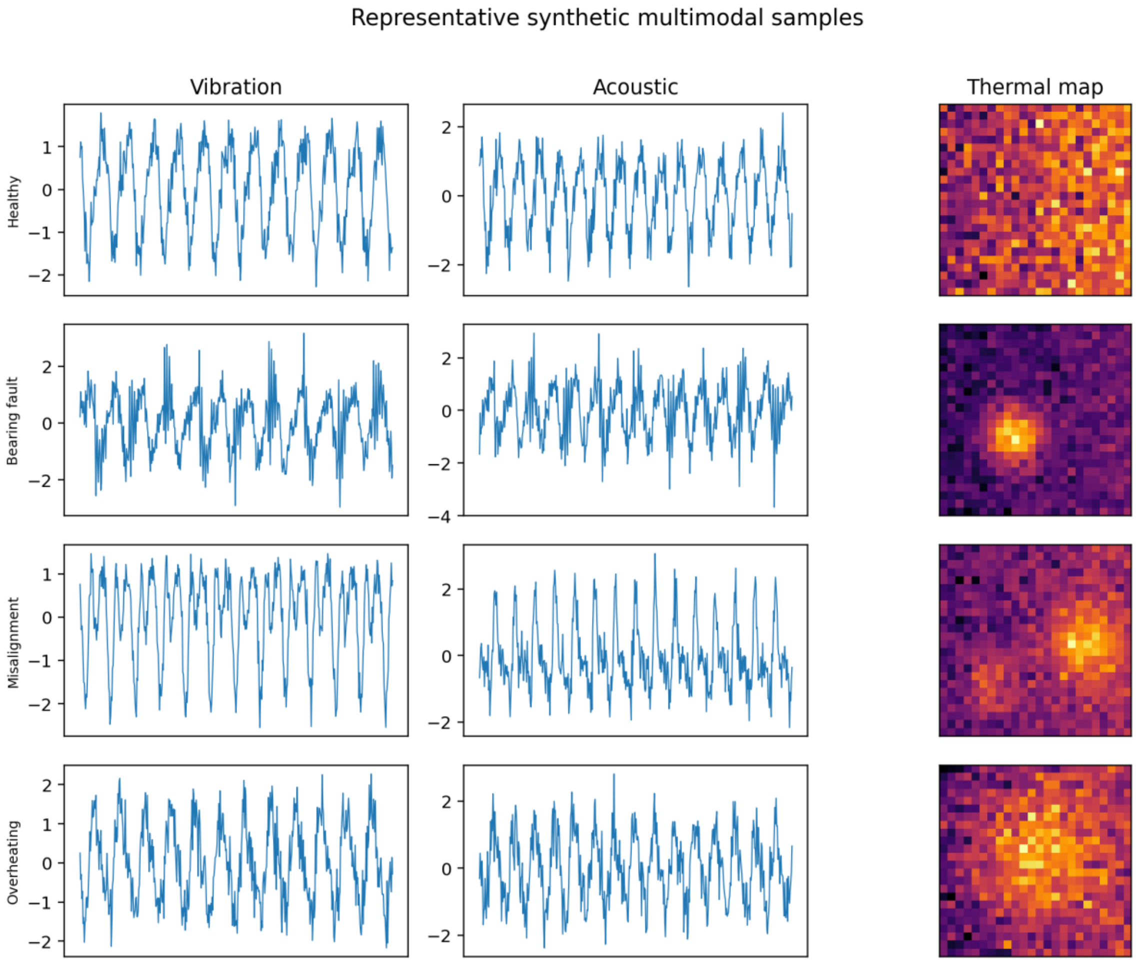

Figure 3. Representative synthetic vibration, acoustic, and thermal observations for the four health states.

4.3 Signal alignment and preprocessing

In a real system, all modalities should be linked to a common asset identifier, operating state, and timestamp. Vibration and acoustic streams may be segmented into short overlapping windows, whereas thermal frames may arrive at a much lower rate. A practical alignment rule is to associate each thermal frame with the nearest waveform window within a maximum time tolerance, while storing the offset as metadata. Alignment quality should be monitored because an incorrect timestamp can create false cross-modal relationships.

Recommended preprocessing includes sensor-health checks, missing-sample detection, anti-alias filtering, detrending, clipping detection, and operating-state tagging. The present experiment used standardized raw waveforms rather than handcrafted statistics. This choice is consistent with the objective of testing end-to-end representation learning, although engineered features such as RMS, kurtosis, crest factor, envelope spectra, mel-spectrograms, and temperature gradients can be added as auxiliary channels in future work.

4.4 Proposed attentive fusion architecture

The architecture contains three modality-specific encoders. The vibration and acoustic branches use one-dimensional convolution, batch normalization, rectified linear units, max pooling, and global average pooling. The thermal branch uses the analogous two-dimensional operations. Each encoder outputs a 32-dimensional feature vector. Separate projection layers map the vectors into a common 32-dimensional latent space.

The three projected vectors are concatenated and passed to a gating network. The gate produces three logits that are converted to non-negative weights summing to one. These weights represent learned sample-wise reliability rather than fixed global importance. The shared representation is the weighted sum of the projected modality vectors. The final classifier receives both the individual modality embeddings and the shared representation, preserving modality-specific evidence while also modeling their interaction.

$$f_v = E_v(x_v), \quad f_a = E_a(x_a), \quad f_t = E_t(x_t)$$

$$w = \mathrm{softmax}(G([f_v ; f_a ; f_t])) = [w_v, w_a, w_t], \quad \text{where } w_v + w_a + w_t = 1$$

$$f_s = w_v f_v + w_a f_a + w_t f_t$$

$$p(y \mid x_v, x_a, x_t) = \mathrm{softmax}(H([f_v ; f_a ; f_t ; f_s]))$$

$$L = -(1/N) \sum_i \sum_c y_{ic} \log(p_{ic})$$

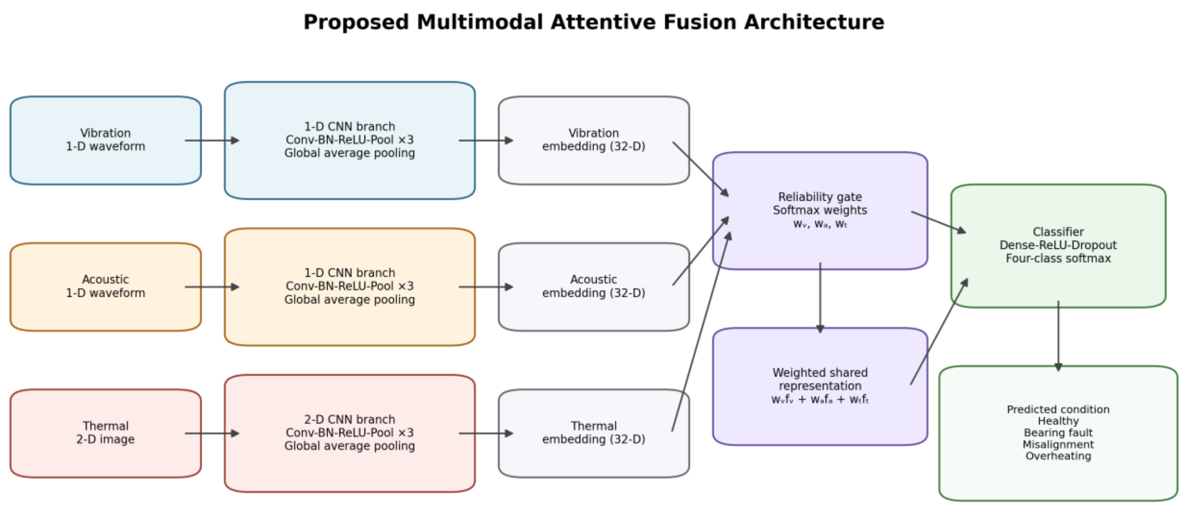

Figure 4. Proposed reliability-aware intermediate-fusion architecture.

Component	Layers	Configuration	Output
Vibration encoder	Conv1D 1->12, Conv1D 12->24, Conv1D 24->32	Kernels 9, 7, 5; pooling and global average pooling	32-D
Acoustic encoder	Same as vibration encoder	Independent parameters	32-D
Thermal encoder	Conv2D 1->12, Conv2D 12->24, Conv2D 24->32	3 x 3 kernels; pooling and global average pooling	32-D
Projection	Three Dense(32->32) + ReLU	Maps modalities to common latent size	3 x 32-D
Reliability gate	Dense(96->48->3) + softmax	Sample-wise modality weights	3 weights
Classifier	Dense(128->64), ReLU, dropout 0.20, Dense(64->4)	Uses individual and shared representations	4 logits

Table 3. Detailed model architecture.

4.5 Baselines, training and evaluation metrics

Three unimodal CNNs used the corresponding encoder followed by a linear four-class classifier. Pairwise results were calculated by averaging the probabilities of two trained unimodal models. Late fusion averaged all three unimodal probability vectors. The proposed model was trained jointly so that its gate and classifier could learn cross-modal interactions.

All networks used cross-entropy loss and the Adam optimizer with a learning rate of 0.001 and weight decay of 0.0001. The batch size was 64. Unimodal baselines were trained for eight epochs, and the proposed model was trained for twelve epochs. The validation checkpoint with the highest validation accuracy was retained. Metrics included accuracy, macro precision, macro recall, macro F1-score, and the confusion matrix. Macro averages were selected because they weight each class equally.

4.6 Robustness and ablation protocol

Robustness was evaluated at waveform-noise levels of 0.45, 0.60, and 0.75, beyond the 0.30 corruption used for the principal test set. Missing-modality tests set one sensor input to zero at inference without retraining. This is a deliberately strict test: production systems should use modality dropout during training, missingness indicators, or a mixture-of-experts design if sensor loss is expected. The ablation therefore measures sensitivity rather than optimized fault tolerance.

5. Results

5.1 Overall model comparison

Model	**Accuracy (%)**	**Macro precision**	**Macro recall**	**Macro F1**
Vibration CNN	76.11	0.8778	0.7611	0.6904
Acoustic CNN	88.89	0.9231	0.8889	0.8848
Thermal CNN	54.44	0.6686	0.5444	0.4578
Vibration + acoustic	79.44	0.8872	0.7944	0.7526
Vibration + thermal	77.78	0.8749	0.7778	0.7203
Acoustic + thermal	88.89	0.9113	0.8889	0.8793
Late probability fusion	82.22	0.8961	0.8222	0.7965
Proposed attentive fusion	**99.44**	**0.9946**	**0.9944**	**0.9944**

Table 4. Principal results on the noise-shifted synthetic test set (n = 180).

The proposed attentive-fusion model achieved 99.44% accuracy and a macro F1-score of 0.9944. It exceeded the strongest unimodal baseline, the acoustic CNN, by 10.56 percentage points. Fixed late probability fusion reached only 82.22%, indicating that joint feature learning and a trained reliability mechanism were more effective than averaging independently trained decisions in this controlled setting.

The thermal-only model produced the lowest score because the noise-shifted evaluation disrupted spatial temperature patterns and because bearing and misalignment classes were designed with partially overlapping hotspot characteristics. Nevertheless, combining thermal information with acoustic evidence improved the pairwise result relative to thermal alone. The results support the central hypothesis that complementary modalities are most useful when their interactions are learned jointly rather than added mechanically.

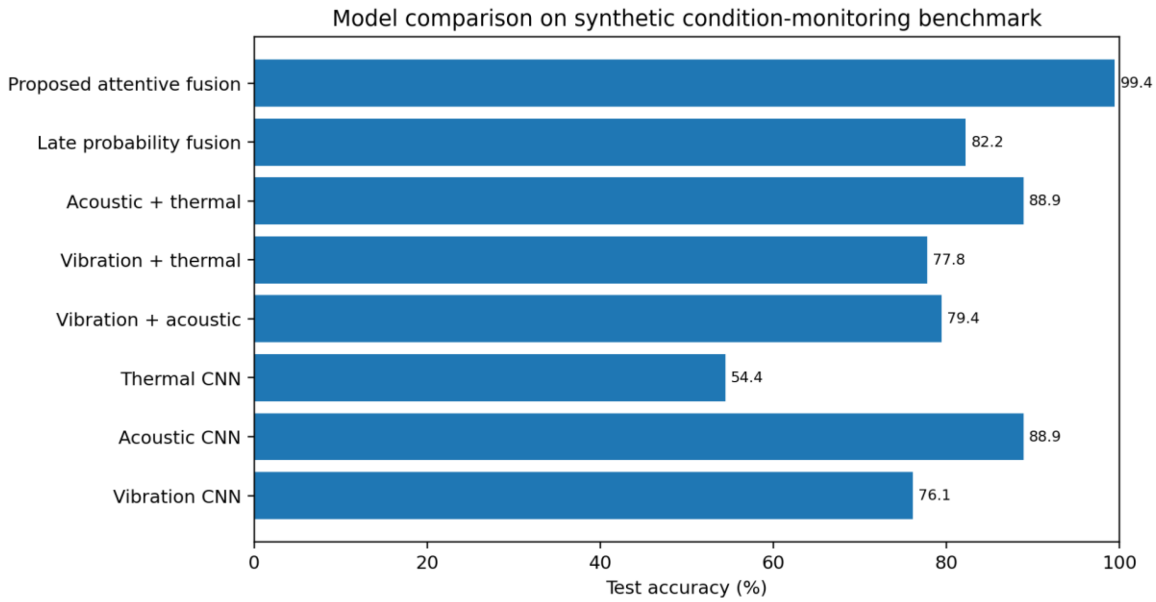

Figure 5. Accuracy comparison across unimodal, pairwise, late-fusion, and proposed models.

5.2 Class-wise performance and confusion analysis

Class	**Precision**	**Recall**	**F1-score**	**Support**
Healthy	0.9783	1.0000	0.9890	45
Bearing fault	1.0000	1.0000	1.0000	45
Misalignment	1.0000	0.9778	0.9888	45
Overheating	1.0000	1.0000	1.0000	45

Table 5. Class-wise performance of the proposed model.

The confusion matrix contained one error: one misalignment observation was classified as healthy. Bearing-fault and overheating observations were classified correctly in all 45 test cases for each class. This pattern is physically plausible within the synthetic design because misalignment shares periodic structure with normal rotation and becomes harder to distinguish when corruption weakens the additional harmonics.

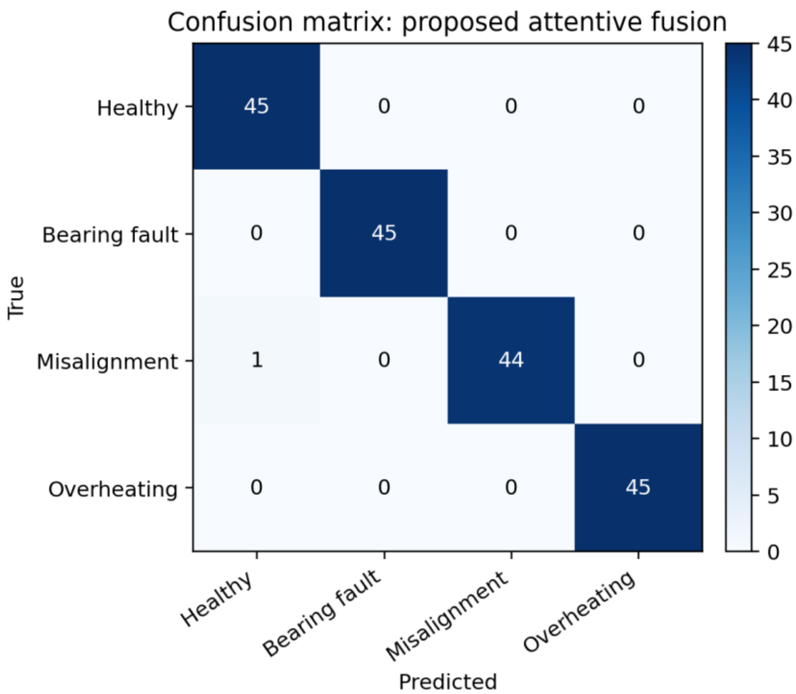

Figure 6. Confusion matrix for the proposed model.

5.3 Training behavior

Training and validation accuracy increased rapidly during the first epochs and stabilized near the optimum. The best validation checkpoint was retained rather than the final epoch by default. Because the synthetic classes contain intentionally structured patterns, convergence is faster than would normally be expected for a heterogeneous industrial dataset. Real applications would require learning-rate schedules, stronger augmentation, cross-validation by asset, and monitoring for overfitting to operating conditions.

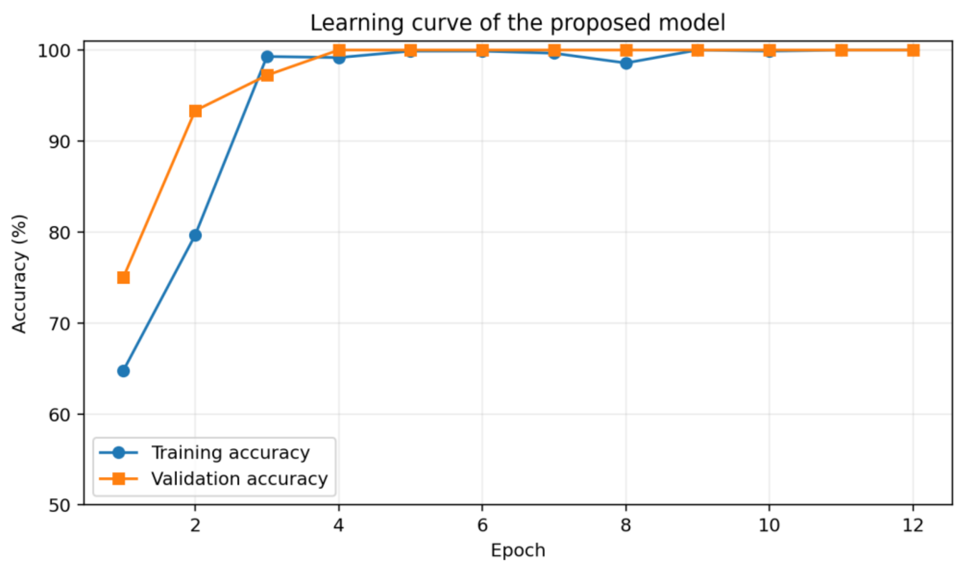

Figure 7. Training and validation accuracy of the proposed attentive-fusion model.

5.4 Noise and missing-modality robustness

Scenario	Accuracy (%)	Macro F1
Main test noise 0.30	99.44	0.9944
Noise 0.45	86.11	0.8513
Noise 0.60	66.67	0.6139
Noise 0.75	41.67	0.3536
Missing vibration	50.56	0.3866
Missing acoustic	96.11	0.9609
Missing thermal	75.00	0.6667

Table 6. Robustness and missing-modality results.

Accuracy decreased from 99.44% at the principal noise level to 86.11% at 0.45, 66.67% at 0.60, and 41.67% at 0.75. This confirms that the model is not invariant to severe corruption. Missing acoustic input caused a relatively small reduction to 96.11%, whereas removing vibration reduced accuracy to 50.56% and removing thermal data reduced it to 75.00%. These values reflect the synthetic mechanisms, not universal modality importance.

The average learned weights on the principal test set were approximately 0.772 for vibration, 0.001 for acoustic, and 0.227 for thermal. The near-zero acoustic gate weight should not be interpreted as proof that acoustic sensing is generally unimportant. The acoustic-only model performed well, but its learned embedding was largely redundant once the jointly trained vibration and thermal features were available. A different machine, sensor placement, or failure mechanism could produce very different weights.

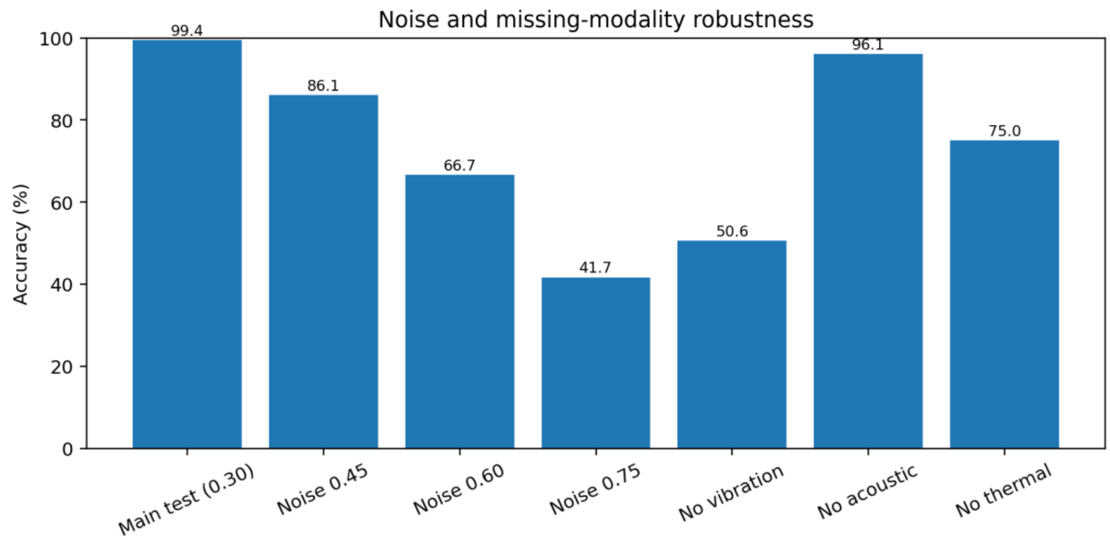

Figure 8. Robustness under increasing corruption and sensor removal.

5.5 Computational and operational interpretation

The model is intentionally compact. Global average pooling reduces parameter count and avoids large fully connected layers. The vibration and acoustic encoders operate directly on short windows, and the 24 x 24 thermal branch is suitable for low-resolution radiometric crops. This architecture can be deployed on an industrial PC or an edge accelerator, but latency must be measured on the target hardware. The model should output class probabilities, a confidence score, sensor-health indicators, and an abstention decision when inputs are outside the validated range.

6. Discussion

6.1 Interpretation of the main findings

The experimental results demonstrate the potential of intermediate fusion when modalities have different physical meanings and noise characteristics. The proposed architecture did not simply average unimodal decisions; it jointly optimized the feature extractors, reliability gate, shared representation, and classifier. This allowed the model to exploit interactions that were not available to independently trained baselines.

The strong result should be interpreted in the context of a structured synthetic benchmark. The classes were generated from separable physical motifs, the training set was balanced, and labels were noise-free. Industrial data are more difficult because operating state, load, speed, mounting, environment, machine age, sensor replacement, maintenance history, and unknown fault combinations create domain shift. Consequently, the reported accuracy is evidence that the architecture can learn the designed relationships, not evidence of field readiness.

The poor performance under missing vibration illustrates a broader lesson: a reliability gate does not automatically guarantee missing-modality robustness. If a model is always trained with all sensors, it may become dependent on the most predictive branch. Modality dropout, sensor-mask tokens, imputation, expert routing, and uncertainty-aware abstention should be included during training when sensors may fail (Neverova et al., 2016).

6.2 Relationship to previous research

The findings are consistent with the general multimodal literature, which argues that fusion is most beneficial when modalities are complementary rather than merely correlated (Baltrusaitis et al., 2019; Ramachandram & Taylor, 2017). The compact 1-D branches follow the efficiency rationale described by Ince et al. (2016) and Kiranyaz et al. (2021), while the thermal branch reflects the spatial nature of infrared evidence documented by Bagavathiappan et al. (2013). The acoustic branch is supported by the availability of industrial sound datasets such as MIMII and ToyADMOS, although the present synthetic waveforms do not reproduce the full complexity of those recordings.

The comparison with late fusion also aligns with earlier multimedia research showing that early and late fusion involve different trade-offs and that no single fusion level is universally optimal (Atrey et al., 2010; Snoek et al., 2005). In safety-critical monitoring, a hybrid architecture may be preferable: feature-level fusion for accuracy, plus independent unimodal heads for redundancy and diagnostic traceability.

6.3 Explainability, uncertainty and human factors

Industrial users require more than a class label. A vibration explanation may highlight waveform segments or frequency bands, an acoustic explanation may identify sound intervals or spectral regions, and a thermal explanation may highlight spatial hotspots. Grad-CAM can support thermal visualization, while SHAP or integrated gradients can summarize feature importance for engineered or latent features (Lundberg & Lee, 2017; Selvaraju et al., 2017). Explanations should be validated with maintenance experts because visually plausible saliency is not necessarily causally correct.

Confidence must be calibrated. A system should abstain when it observes an unknown operating state, sensor saturation, missing synchronization, or a probability distribution that is inconsistent with the validation data. Deep ensembles, temperature scaling, conformal prediction, or Bayesian approximations can be considered. False negatives and false positives should be weighted according to asset criticality, safety consequences, downtime cost, and inspection burden.

Human-machine interaction is central. The model should support, not replace, maintenance judgment. Alerts should include the suspected condition, confidence, recent trend, affected sensors, recommended inspection, and evidence. Alarm flooding can erode trust; therefore, alert persistence, hysteresis, severity levels, and work-order feedback are necessary.

6.4 Threats to validity and limitations

Synthetic data do not capture the full physics, sensor transfer paths, non-stationary operating conditions, compound faults, or environmental interference of industrial assets.

The benchmark is balanced and uses perfect labels. Real datasets are often imbalanced, weakly labeled, and dominated by healthy operation.

The thermal maps are low-resolution standardized arrays rather than calibrated radiometric images with emissivity and ambient corrections.

The waveform corruption model is Gaussian, whereas field noise may be colored, impulsive, periodic, reverberant, or generated by neighboring equipment.

The missing-modality test does not retrain the model with modality dropout and therefore represents a strict sensitivity test rather than an optimized robust system.

No independent public tri-modal dataset was available for direct validation, and no cross-asset or cross-site generalization experiment was performed.

Inference latency, energy consumption, calibration, and long-term drift were not measured on industrial hardware.

7. Industrial Deployment Blueprint

7.1 Sensor and acquisition layer

Use industrial accelerometers with suitable bandwidth and mounting, microphones or acoustic sensors positioned to maximize asset signal relative to background noise, and radiometric thermal cameras with controlled field of view. Store sensor serial number, calibration date, location, axis, sampling rate, gain, emissivity setting, and firmware version.

Synchronize devices through a common clock or timestamp service. Measure alignment error and reject fused observations when the error exceeds a validated threshold. Triggered acquisition can reduce data volume, but continuous low-rate health features should remain available for trend analysis.

7.2 Data engineering and model lifecycle

Organize data by asset, operating state, load, speed, maintenance event, and fault confirmation. Prevent leakage by splitting evaluation data by asset or time period rather than randomly mixing adjacent windows. Version the raw data, preprocessing code, labels, model, thresholds, and deployment configuration.

Begin with a shadow-mode pilot in which the model produces alerts without controlling maintenance. Compare predictions with inspections and work orders. Establish baseline false-alarm rates, missed-fault risk, and lead time. Retrain only after documented label review and change control.

7.3 Edge-cloud architecture and cybersecurity

A practical architecture can perform waveform segmentation and inference at the edge while sending summarized features, selected waveforms, thermal crops, and alerts to a central platform. Edge inference reduces bandwidth and latency, while centralized storage supports fleet learning and auditability.

Condition-monitoring devices should use authenticated communication, encrypted transport, secure boot where available, network segmentation, least-privilege access, signed model updates, and event logging. Sensor spoofing, timestamp manipulation, and unauthorized model replacement are relevant risks because they can suppress or create maintenance alarms.

Control area	**Key checks**	**Recommended action**
Data quality	Missing samples, clipping, drift, time offset, thermal saturation	Reject or flag invalid windows before inference
Model quality	Accuracy, macro F1, calibration, false-alarm rate, lead time	Evaluate by asset, fault type, load, and season
Robustness	Noise, missing sensor, domain shift, unknown fault	Abstain and request inspection when outside limits
Operations	Alert latency, acknowledgment time, work-order closure	Integrate with CMMS and escalation rules

Governance	Model version, approval, change log, rollback	Controlled deployment and audit trail
Cybersecurity	Authentication, encryption, secure update, network segmentation	Protect sensing and decision chain

Table 7. Minimum deployment controls for an industrial multimodal monitoring system.

8. Conclusion and Future Work

This paper presented a complete multimodal deep-learning framework for intelligent condition monitoring using vibration, acoustic, and thermal data. Modality-specific CNNs preserved the distinct structure of waveform and image inputs, while a learned reliability gate produced sample-wise fusion weights and a shared representation. On a transparent synthetic four-condition benchmark with a noise-shifted test set, the proposed model achieved 99.44% accuracy and outperformed unimodal and fixed late-fusion baselines.

The most important conclusion is methodological rather than numerical: heterogeneous sensor information should be aligned, represented, fused, and validated as part of one monitored system. Reliability-aware fusion can improve classification, but it must be paired with sensor-health checks, uncertainty estimation, missing-modality training, explainability, and maintenance workflow integration. The sharp degradation under severe corruption and sensor removal shows why robust evaluation is essential.

Future work should validate the architecture on synchronized real-machine data from multiple assets and sites; include variable speed, load, and compound faults; use self-supervised pretraining for scarce labels; add modality dropout and missingness indicators; investigate cross-modal transformers and mixture-of-experts models; calibrate uncertainty and abstention; compare raw-waveform and time-frequency representations; measure edge latency and energy use; and evaluate whether multimodal alerts improve maintenance lead time and economic outcomes. Public release of synchronized tri-modal datasets would substantially accelerate progress in this field.

Declarations

Funding: No external funding was used for this proof-of-concept manuscript.

Conflict of interest: The author declares no conflict of interest.

Data availability: The reported results were generated from synthetic data using the configuration described in the paper. The data-generation and model scripts should be archived with any submitted version.

Ethics statement: The study used no human participants, animals, or personal data.

Author contribution: [Insert author contribution statement before submission].

References

1. Abdeljaber, O., Avci, O., Kiranyaz, S., Boashash, B., Sodano, H., & Inman, D. J. (2017). Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. Journal of Sound and Vibration, 388, 154-170. https://doi.org/10.1016/j.jsv.2016.10.043
2. Antoni, J. (2006). The spectral kurtosis: A useful tool for characterising non-stationary signals. Mechanical Systems and Signal Processing, 20(2), 282-307. https://doi.org/10.1016/j.ymssp.2004.09.001
3. Arevalo, J., Solorio, T., Montes-y-Gomez, M., & Gonzalez, F. A. (2017). Gated multimodal units for information fusion. In ICLR 2017 Workshop Track Proceedings.
4. Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. Multimedia Systems, 16, 345-379. https://doi.org/10.1007/s00530-010-0182-0
5. Bagavathiappan, S., Lahiri, B. B., Saravanan, T., Philip, J., & Jayakumar, T. (2013). Infrared thermography for condition monitoring - A review. Infrared Physics & Technology, 60, 35-55. https://doi.org/10.1016/j.infrared.2013.03.006
6. Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443. https://doi.org/10.1109/TPAMI.2018.2798607
7. Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review. Data Mining and Knowledge Discovery, 33, 917-963. https://doi.org/10.1007/s10618-019-00619-1
8. Glowacz, A., & Glowacz, Z. (2017). Diagnosis of the three-phase induction motor using thermal imaging. Infrared Physics & Technology, 81, 7-16. https://doi.org/10.1016/j.infrared.2016.12.003
9. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

10. Hall, D. L., & Llinas, J. (1997). An introduction to multisensor data fusion. Proceedings of the IEEE, 85(1), 6-23. https://doi.org/10.1109/5.554205
11. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90
12. Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017). CNN architectures for large-scale audio classification. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 131-135). https://doi.org/10.1109/ICASSP.2017.7952132
13. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735
14. Huda, A. S. N., & Taib, S. (2013). Application of infrared thermography for predictive/preventive maintenance of thermal defect in electrical equipment. Applied Thermal Engineering, 61(2), 220-227. https://doi.org/10.1016/j.applthermaleng.2013.07.028
15. Ince, T., Kiranyaz, S., Eren, L., Askar, M., & Gabbouj, M. (2016). Real-time motor fault detection by 1-D convolutional neural networks. IEEE Transactions on Industrial Electronics, 63(11), 7067-7075. https://doi.org/10.1109/TIE.2016.2582729
16. Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning (pp. 448-456). PMLR.
17. Janssens, O., Slavkovikj, V., Vervisch, B., Stockman, K., Loccufier, M., Verstockt, S., Van de Walle, R., & Van Hoecke, S. (2016). Convolutional neural network based fault detection for rotating machinery. Journal of Sound and Vibration, 377, 331-345. https://doi.org/10.1016/j.jsv.2016.05.027
18. Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing, 20(7), 1483-1510. https://doi.org/10.1016/j.ymssp.2005.09.012
19. Khan, S., & Yairi, T. (2018). A review on the application of deep learning in system health management. Mechanical Systems and Signal Processing, 107, 241-265. https://doi.org/10.1016/j.ymssp.2017.11.024
20. Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. Information Fusion, 14(1), 28-44. https://doi.org/10.1016/j.inffus.2011.08.001
21. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In International Conference on Learning Representations.
22. Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. Mechanical Systems and Signal Processing, 151, 107398. https://doi.org/10.1016/j.ymssp.2020.107398
23. Koizumi, Y., Saito, S., Uematsu, H., Harada, N., & Imoto, K. (2019). ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection. In 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (pp. 313-317). https://doi.org/10.1109/WASPAA.2019.8937164
24. Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. Mechanical Systems and Signal Processing, 104, 799-834. https://doi.org/10.1016/j.ymssp.2017.11.016
25. Lessmeier, C., Kimotho, J. K., Zimmer, D., & Sextro, W. (2016). Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. PHM Society European Conference, 3(1). https://doi.org/10.36001/phme.2016.v3i1.1577
26. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30.
27. Neverova, N., Wolf, C., Taylor, G. W., & Nebout, F. (2016). ModDrop: Adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(8), 1692-1706. https://doi.org/10.1109/TPAMI.2015.2461544
28. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning (pp. 689-696).
29. Piczak, K. J. (2015). Environmental sound classification with convolutional neural networks. In 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing. https://doi.org/10.1109/MLSP.2015.7324337
30. Purohit, H., Tanabe, R., Ichige, T., Endo, T., Nikaido, Y., Suefusa, K., & Kawaguchi, Y. (2019). MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (pp. 209-213).
31. Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine, 34(6), 96-108. https://doi.org/10.1109/MSP.2017.2738401
32. Randall, R. B., & Antoni, J. (2011). Rolling element bearing diagnostics - A tutorial. Mechanical Systems and Signal Processing, 25(2), 485-520. https://doi.org/10.1016/j.ymssp.2010.07.017
33. Sawalhi, N., Randall, R. B., & Endo, H. (2007). The enhancement of fault detection and diagnosis in rolling element bearings using minimum entropy deconvolution combined with spectral kurtosis. Mechanical Systems and Signal Processing, 21(6), 2616-2633. https://doi.org/10.1016/j.ymssp.2006.12.002

34. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618-626). https://doi.org/10.1109/ICCV.2017.74
35. Smith, W. A., & Randall, R. B. (2015). Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. Mechanical Systems and Signal Processing, 64-65, 100-131. https://doi.org/10.1016/j.ymssp.2015.04.021
36. Snoek, C. G. M., Worring, M., & Smeulders, A. W. M. (2005). Early versus late fusion in semantic video analysis. In Proceedings of the 13th Annual ACM International Conference on Multimedia (pp. 399-402). https://doi.org/10.1145/1101149.1101236
37. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15, 1929-1958.
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30.
39. Wang, Z., Yan, W., & Oates, T. (2017). Time series classification from scratch with deep neural networks: A strong baseline. In 2017 International Joint Conference on Neural Networks (pp. 1578-1585). https://doi.org/10.1109/IJCNN.2017.7966039
40. Wen, L., Li, X., Gao, L., & Zhang, Y. (2018). A new convolutional neural network-based data-driven fault diagnosis method. IEEE Transactions on Industrial Electronics, 65(7), 5990-5998. https://doi.org/10.1109/TIE.2017.2774777
41. Yang, X., Fang, C., Liao, Y., Yang, J., Gryllias, K., & Chronopoulos, D. (2026). Deep generative models in condition and structural health monitoring: Opportunities, limitations and future outlook. Mechanical Systems and Signal Processing, 253, 114277. https://doi.org/10.1016/j.ymssp.2026.114277
42. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1103-1114). https://doi.org/10.18653/v1/D17-1115
43. Zhang, W., Peng, G., Li, C., Chen, Y., & Zhang, Z. (2018). A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. Mechanical Systems and Signal Processing, 100, 439-453. https://doi.org/10.1016/j.ymssp.2017.06.022
44. Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing, 115, 213-237. https://doi.org/10.1016/j.ymssp.2018.05.050