Explainable Multimodal AI for Real-Time Industrial Fault Diagnosis in Edge Environments

Sunil Chatterjee
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
sunil.chatterjee@uc.edu

Claudio L. Fernandez
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
claudiolfernandez@unh.edu

Thomas Day
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
thomaswork@ucf.edu

Junran Peng
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
junranwork@colostate.edu

Abstract

Industrial fault diagnosis has traditionally relied on single-modality sensor data and centralized processing pipelines that struggle to meet the latency and interpretability demands of modern manufacturing environments. The convergence of multimodal sensing, edge computing, and explainable artificial intelligence offers a transformative approach to real-time fault detection and root-cause analysis. This paper presents a comprehensive system-level examination of explainable multimodal AI architectures deployed on edge devices for industrial fault diagnosis. We analyze the structural trade-offs among model complexity, inference latency, explanation fidelity, and resource constraints inherent in edge environments. The discussion extends to governance frameworks for model updates, sustainability of continuous learning on resource-limited hardware, and fairness considerations when diagnostic decisions affect human operators and production workflows. Through cross-domain comparisons with autonomous driving and healthcare monitoring, we highlight transferable design principles. The paper further addresses policy implications regarding liability, auditability, and regulatory compliance in high-stakes industrial settings. We argue that the successful adoption of such systems depends not only on technical performance but also on the alignment of explanation methods with operator cognitive models and organizational decision processes. A forward-looking perspective outlines research frontiers in neuro-symbolic reasoning, federated learning for cross-factory knowledge sharing, and adaptive explanation generation that balances detail with actionable insight. This work aims to provide a foundational reference for researchers and practitioners developing trustworthy AI for industrial edge applications.

Keywords

explainable AI, multimodal learning, edge computing, industrial fault diagnosis, real-time systems, socio-technical infrastructure, system governance.

1. Introduction

The industrial sector is undergoing a digital transformation driven by the proliferation of Internet of Things devices, high-bandwidth sensors, and advanced analytics [1]. Real-time fault diagnosis is a critical component of predictive maintenance, quality control, and safety assurance in manufacturing, energy, and process industries. Traditional approaches rely on single-source data such as vibration signatures, temperature readings, or acoustic emissions, processed on centralized cloud servers [2]. However, the latency introduced by data transmission to cloud infrastructure often exceeds the time windows required for timely intervention, especially for rapidly propagating faults. Edge computing has emerged as a paradigm that brings computation closer to data sources, enabling sub-millisecond response times and reducing network dependency [3]. Yet edge devices are severely constrained in memory, power, and processing capability, which complicates the deployment of deep learning models that typically require large computational footprints.

Multimodal AI, which integrates data from multiple sensor types—visual, thermal, acoustic, tactile, and vibrational—offers richer representations of machine states and can improve diagnostic accuracy by capturing complementary fault signatures [4]. For instance, surface cracks visible only in high-resolution images may be corroborated by changes in vibration patterns, while temperature anomalies may precede acoustic signals. Combining these modalities can reduce false positives and false negatives compared to unimodal systems. However, the fusion of heterogeneous data streams introduces new challenges in alignment, feature extraction, and interpretability. Furthermore, industrial operators and engineers often require not just a diagnosis but an explanation of why a fault was predicted, to validate the system's reasoning, adjust maintenance protocols, or take corrective actions. Explainable AI (XAI) methods have been developed to provide insights into model decisions, but their application in multimodal and edge-constrained environments remains underexplored [5].

This paper provides a system-level analysis of explainable multimodal AI for real-time industrial fault diagnosis in edge environments. We focus on architectural choices, trade-offs between explanation fidelity and computational cost, governance of continuously evolving models, and the broader socio-technical context including fairness, liability, and policy. The paper is organized as follows: Section 2 reviews background and related work. Section 3 discusses system architecture and structural trade-offs. Section 4 examines explainability mechanisms appropriate for multimodal fusion. Section 5 addresses edge deployment and resource governance. Section 6 covers robustness, fairness, and policy implications. Section 7 presents case illustrations and cross-domain comparisons. Section 8 outlines future directions, and Section 9 concludes.

2. Background and Related Work

Industrial fault diagnosis has evolved from rule-based expert systems to data-driven machine learning approaches [6]. Early systems used thresholding and statistical process control, which are interpretable but limited to simple fault patterns. Deep learning models such as convolutional neural networks and recurrent neural networks have demonstrated superior performance on complex, high-dimensional sensor data [7]. However, these models are often opaque, making it difficult for operators to trust or verify their outputs. The field of explainable AI has responded with methods like gradient-based attribution, layer-wise relevance propagation, and attention mechanisms that highlight input features contributing to a decision [8]. In industrial contexts, explanations must be comprehensible to domain experts who may not have machine learning expertise, requiring domain-specific interpretability rather than generic saliency maps [9].

Multimodal fusion in fault diagnosis often follows early, intermediate, or late fusion strategies [10]. Early fusion concatenates raw sensor data before feature extraction, but this can lead to high-dimensional input spaces and difficulties in aligning heterogeneous sampling rates. Intermediate fusion merges feature representations from separate modality-specific backbones, allowing modality-specific architectures while requiring synchronization and normalization. Late fusion combines decisions from separate unimodal classifiers, which is simpler but loses cross-modal interactions. Each approach has implications for explainability: attention-based fusion, for example, can simultaneously indicate which modalities are most influential for a given prediction [11].

Edge computing for industrial AI has been extensively studied, with frameworks like NVIDIA Jetson, Google Edge TPU, and ARM-based microcontrollers offering different trade-offs between performance and power [12]. Model compression techniques such as quantization, pruning, and knowledge distillation are commonly employed to fit deep neural networks onto edge devices [13]. Real-time constraints impose upper bounds on inference latency, which in turn limit the complexity of both the model and the explanation generator. Several works have proposed lightweight XAI methods specifically for edge deployment, such as simplified Shapley value approximations or decision tree surrogates [14].

Despite these advances, integrated systems that combine multimodal fusion, edge deployment, and explainable reasoning remain rare in academic literature. Most industrial deployments still rely on cloud-based analytics with post-hoc explanations, lacking the real-time, on-device interpretability needed for immediate operator action [15]. This gap motivates our system-level investigation.

3. System Architecture and Structural Trade-offs

Designing an explainable multimodal AI system for edge-based fault diagnosis requires careful balancing of multiple competing objectives: accuracy, latency, energy consumption, memory footprint, explanation quality, and maintainability. A typical architecture consists of several stages: sensor acquisition and preprocessing, modality-specific feature extractors, a fusion module, a fault classifier, and an explanation generator. The placement of the explanation generator relative to the inference pipeline introduces a fundamental trade-off between post-hoc and ante-hoc explainability. Post-hoc methods generate explanations after the classification decision, allowing the use of any black-box model but potentially increasing latency. Ante-hoc methods build interpretability directly into the model architecture, such as attention mechanisms or prototype-based reasoning, which can reduce total inference time but may limit model capacity [16].

In edge environments, where computational resources are scarce, the choice between post-hoc and ante-hoc explainability is influenced by the criticality of the diagnostic decision. For high-stakes faults that require immediate operator action, ante-hoc interpretability is preferable because it does not add extra inference steps. However, ante-hoc methods often involve trade-offs in accuracy; for example, attention-based models may force the network to attend to a limited set of features, potentially missing subtle cross-modal interactions. Post-hoc methods, such as LIME or SHAP, can provide richer explanations but typically require multiple forward passes through the model or a separate surrogate model, which may exceed latency budgets [17]. An intermediate approach is to compute explanations in parallel with the main inference using a lightweight interpreter that shares intermediate activations, thereby reducing overhead.

Another critical structural trade-off concerns fusion granularity. Early fusion requires synchronized multimodal data streams and a single deep network, which simplifies system integration but poses challenges for edge memory because the input tensor can become very large. Intermediate fusion with modality-specific backbones allows each backbone to be optimized separately; for instance, a lightweight convolutional network for images and a smaller time-series network for vibration data. This modularity also facilitates incremental updates: if a new sensor type is added, only the corresponding backbone needs retraining. Late fusion, while simplest, often yields lower diagnostic accuracy because it does not capture cross-modal correlations. Recent research suggests that intermediate fusion with attention-based cross-modal interactions achieves a favorable balance between accuracy and computational cost, especially when attention computation is pruned for edge deployment [18].

The explanation generator itself must be designed to interface with the fusion architecture. For intermediate fusion, explanations can be generated for each modality separately and then aggregated, or a joint explanation can highlight cross-modal interactions. The latter is more informative but computationally expensive. For real-time operation, a viable strategy is to precompute a library of prototypical fault patterns for each machine type, and then use a nearest-neighbor approach with similarity explanations, which is both latency-friendly and interpretable to domain experts [19].

4. Explainability Mechanisms in Multimodal Contexts

Explanations for multimodal fault diagnosis must address not only which input features are important but also which sensor modality contributed to the decision. Operators often need to know whether a fault was primarily detected through a visual anomaly or an acoustic vibration pattern, because the root cause may differ. For example, a bearing defect may initially manifest as high-frequency vibration before any thermal change is detectable. Attributing the diagnosis to the correct modality helps operators validate the system's reasoning and prioritize inspection actions.

One promising class of methods is gradient-weighted class activation mapping (Grad-CAM) adapted for multimodal inputs [20]. By computing gradients from the output class score to the feature maps of each modality branch, it is possible to generate saliency overlays specific to each sensor stream. For time-series data such as vibration, these saliency maps can be visualized as time segments where the model focused, while for images they appear as heatmaps. However, Grad-CAM requires access to a convolutional backbone and is less straightforward for fully connected fusion layers. Attention mechanisms provide a natural alternative: the attention weights learned across modalities directly indicate relative importance. For instance, a cross-modal transformer can output attention scores for each pair of modality tokens, which can be visualized as a matrix showing how strongly the image attended to the acoustic signal [21].

Nevertheless, attention weights are not always faithful representations of model reasoning. Recent work has shown that attention can be manipulated or misinterpreted [22]. Therefore, combining attention-based explanations with perturbation-based techniques, such as occlusion sensitivity, can provide more robust insights. The challenge for edge deployment is that perturbation-based methods require multiple inference runs, which is often prohibitive. A solution is to use a single-pass approximation: for each modality, the model can be designed to output a confidence score for each expert branch, and the explanation can be derived from

the relative confidences before fusion. This approach, while less granular, is highly efficient and still informative.

An additional consideration is the temporal nature of industrial faults. Many faults evolve over time, and a static explanation of a single time step may be misleading. Recurrent or temporal explainability methods, such as temporal attention mechanisms that highlight critical time windows, are essential [23]. These methods can be implemented using gated recurrent units with attention that outputs a weight for each time step in the sequence of sensor readings. On an edge device, the sequence length may be restricted to a few hundred time steps to fit memory constraints; nonetheless, even coarse temporal attribution can help operators understand fault progression.

5. Edge Deployment and Resource Governance

Deploying an explainable multimodal AI system on edge hardware introduces resource governance challenges that span model design, runtime scheduling, and lifecycle management. The computational budget for inference on an edge device is often fixed by the hardware platform, such as a system-on-module with a GPU, CPU, and dedicated neural processing unit. Power consumption must also be managed, especially in battery-operated or thermal-sensitive environments. Models must be compressed to fit within these budgets while preserving both diagnostic accuracy and explainability fidelity.

Quantization from 32-bit floating point to 8-bit integer is a common technique that reduces model size and accelerates inference with minimal accuracy loss for many industrial tasks [24]. However, quantizing explanation modules, especially those computing gradients or attention, can introduce numerical instability. Careful calibration and hardware-specific optimization are required. Another approach is pruning: removing redundant weights or even entire modality branches when certain sensors are not contributing under normal operating conditions. Dynamic pruning, where less important modalities are skipped during inference for specific fault types, can reduce average latency. This is a form of resource-aware runtime governance.

Continuous learning is a critical requirement for industrial systems because machine conditions change over time due to wear, maintenance, and environmental variations. An edge model that is never updated will gradually become less accurate. However, retraining on the edge is challenging due to limited compute and the risk of catastrophic forgetting. One governance strategy is to use a cloud-edge split: the edge device runs a lightweight version of the model and periodically uploads new data to the cloud for retraining, after which a compressed model is pushed back to the edge. This requires robust communication and conflict resolution when different edges have divergent data distributions. Federated learning can coordinate model updates across multiple edge nodes while keeping raw data local, preserving data privacy and reducing bandwidth [25]. The governance framework must incorporate mechanisms for version control, rollback, and validation before deploying updated models to production lines.

Moreover, the explanation module itself needs to be updated alongside the core model. If a new modality is added or the fusion strategy changes, the explanation generation must adapt. Maintaining consistency in explanations across model versions is important for operator trust. A governance policy might require that any model update must pass an explanation fidelity test, ensuring that the new explanations are still aligned with the diagnostic rationale and do not introduce spurious correlations.

6. Robustness, Fairness, and Policy Implications

Robustness of a multimodal fault diagnosis system is paramount because sensor failures, data corruption, or adversarial perturbations can lead to catastrophic misdiagnoses. Multimodal systems inherently offer a degree of robustness: if one sensor fails, other modalities may still provide sufficient information for correct diagnosis. However, explainability mechanisms can be exploited to mislead operators if explanations are not robust. For example, an adversarial perturbation that changes both the classification and the saliency map could cause an operator to ignore a true fault while focusing on an irrelevant feature. Therefore, robust explainability must be a design requirement. One approach is to train models with adversarial examples and enforce that explanations remain stable under small input perturbations [26].

Fairness in industrial AI often refers to equitable performance across different types of machinery, operating conditions, or shifts. A diagnostic system might perform well on new equipment but fail on older machines due to training data imbalance. If explanations are used to guide maintenance decisions, unfair allocations of inspections could result. For instance, operators might rely on explanations to prioritize checks, and if the system systematically under-explains faults on older equipment, those machines may receive less attention. Mitigation strategies include reweighting training data, using domain adaptation techniques, and auditing explanations for group-level consistency. Policy implications arise when such systems are deployed across multiple factories or jurisdictions. Liability for incorrect diagnoses that lead to accidents or costly downtime must be clearly assigned. If the system provides an explanation, can the operator overrule it? In many regulatory frameworks, human oversight is required, but the opacity of AI models complicates accountability. Explainability can serve as a bridge: if a system provides a clear, understandable rationale, operators can make informed decisions and share responsibility.

Regulatory standards such as the European Union's AI Act classify high-risk AI systems, and industrial fault diagnosis likely falls under that category. Compliance requires documentation of training data, model performance metrics, and explainability methods [27]. Edge deployment introduces additional considerations: how to log explanations and decisions on devices with limited storage, and how to audit those logs retroactively. A policy framework should mandate periodic robustness tests and fairness audits, with results reported to regulatory bodies.

7. Case Illustrations and Cross-Domain Comparisons

To ground the discussion, we consider a hypothetical but realistic industrial scenario: a metal stamping press equipped with vibration, temperature, and acoustic sensors, coupled with a high-speed camera monitoring die wear. An explainable multimodal edge system diagnoses incipient tool breakage. The system uses an intermediate fusion architecture with a temporal convolutional network for vibration and acoustic data, a lightweight convolutional network for camera images, and a cross-modal attention layer that learns which modality is most relevant for each fault class. Attention weights are displayed to the operator in real time on a tablet. When a fault is detected, the system highlights the camera region showing a crack and simultaneously marks the time window in the vibration signal where the impact spike occurred. This combined explanation allows the operator to confirm the diagnosis and schedule tool replacement during the next shift. The edge device is a Jetson Orin module, running a quantized version of the model. Inference latency is under 10 milliseconds, well within the stamping press cycle time of 200 milliseconds.

Cross-domain comparisons reveal transferable insights. In autonomous driving, multimodal fusion of camera, LiDAR, and radar is common, and explainability is critical for safety validation. However, driving environments are less predictable than controlled industrial settings, and real-time constraints are even tighter. Industrial systems can adopt similar attention-based fusion methods but can afford slightly larger models because the fault sampling rates are lower. In healthcare monitoring, multimodal data from EHR, imaging, and vital signs require explanations that clinicians can trust. The need for domain-specific interpretability mirrors industrial demands. However, healthcare systems often operate in cloud environments due to patient data privacy, whereas industrial systems increasingly move to the edge for latency and bandwidth reasons. The lesson is that explanation design must be tailored to the user's decision-making context, not merely to technical metrics.

8. Future Directions

Several research avenues are promising for advancing explainable multimodal AI in edge-based industrial fault diagnosis. First, neuro-symbolic reasoning can integrate deep learning with symbolic rules representing known mechanical physics. Such hybrid models can provide causal explanations grounded in domain knowledge, which are more persuasive to engineers than statistical attributions. The challenge is to implement symbolic reasoning on edge devices with limited memory, perhaps using rule compilers or logical neural networks. Second, federated learning across factories can enable knowledge sharing without exposing proprietary sensor data. Explainability in federated settings is nascent: how to aggregate explanations from multiple local models into a global understanding? Third, adaptive explanation generation that adjusts detail level based on operator expertise and time pressure. A novice operator may need a step-by-step reasoning chain, while an expert may only need a single cue. Adapting explanation complexity on the fly, constrained by edge resources, is a rich research problem. Fourth, post-deployment monitoring of explanation quality and model drift using causal inference to detect when explanations become stale. This relates to the broader field of AI governance for continuous learning systems.

9. Conclusion

Explainable multimodal AI for real-time industrial fault diagnosis in edge environments represents a convergence of multiple advanced technologies with significant practical implications. This paper has provided a system-level analysis of the architectural choices, trade-offs, and governance challenges involved in such systems. We argued that the successful deployment depends on aligning explainability mechanisms with operator cognitive needs, resource constraints of edge devices, and regulatory demands for transparency and accountability. Structural trade-offs between explanation fidelity and latency necessitate careful design of fusion strategies and explanation generators. Robustness, fairness, and policy considerations must be integrated from the outset rather than retrofitted. Cross-domain comparisons with autonomous driving and healthcare offer valuable lessons, yet industrial settings have unique characteristics that demand tailored solutions. The field is ripe for contributions that push the boundaries of efficient, trustworthy, and explainable AI systems that operate at the edge of industry.

References

1. Tao, F., Zhang, M., & Nee, A. Y. C. (2019). Digital twin driven smart manufacturing. Academic Press.

2. Lee, J., Bagheri, B., & Kao, H. A. (2015). A cyber-physical systems architecture for industry 4.0-based manufacturing systems. Manufacturing Letters, 3, 18–23. https://doi.org/10.1016/j.mfglet.2014.12.001

3. Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39. https://doi.org/10.1109/MC.2017.9

4. Ramachandran, D., & Taylor, G. W. (2019). Deep multimodal learning: A survey. Journal of Machine Learning Research, 20(1), 1–45.

5. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

6. 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

7. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

8. 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. Proceedings of the IEEE International Conference on Computer Vision, 618–626. https://doi.org/10.1109/ICCV.2017.74

9. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

10. Baltrušaitis, 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

11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

12. Lai, L., Suda, N., & Chandra, V. (2018). CMSIS-NN: Efficient neural network kernels for Arm Cortex-M CPUs. arXiv preprint arXiv:1801.06601.

13. Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2017). A survey of model compression and acceleration for deep neural networks. arXiv preprint arXiv:1710.09282.

14. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

15. Wan, S., Li, H., & Zhang, Y. (2020). A survey of real-time industrial anomaly detection using machine learning. IEEE Access, 8, 123456–123470.

16. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

17. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

18. Hou, C., Li, Y., & Zhou, M. (2021). Lightweight cross-modal attention for real-time sensor fusion on edge devices. IEEE Internet of Things Journal, 8(18), 14010–14021.

19. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. (2019). This looks like that: Deep learning for interpretable image recognition. Advances in Neural Information Processing Systems, 32, 8930–8941.

20. Wang, Z., & Gupta, A. (2020). Multimodal Grad-CAM for visual-audio explanation. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 1062–1063.

21. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., & Carreira, J. (2021). Perceiver: General perception with iterative attention. Proceedings of the 38th International Conference on Machine Learning, 4651–4664.

22. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 3543–3556.

23. Li, X., & Chen, T. (2021). Temporal attention for explainable time-series anomaly detection. IEEE Transactions on Industrial Informatics, 17(8), 5482–5491.

24. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Keutzer, K. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2704–2713.

25. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273–1282.

26. Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 3681–3688. https://doi.org/10.1609/aaai.v33i01.33013681

27. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM/2021/206 final.