Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Toward GAN-Based Multimodal Novelty Detection for Industrial Fault Taxonomy Expansion: A Feasibility Study

Synthetic-Data Proof-of-Concept for a Grounded, Human-in-the-Loop Failure-Mode Taxonomy Expansion Pipeline

Dinesh Surisetti

Abstract - Existing industrial condition-monitoring systems classify equipment faults against a fixed, pre-defined taxonomy, or mine historical maintenance text records after the fact; neither approach is designed to detect and describe a failure signature absent from the existing taxonomy at the moment it occurs, directly from raw multimodal sensor data. We propose an architecture that fuses synchronized video, audio, and sensor telemetry into a per-cycle embedding, scores it against a model of known-normal operation to obtain a per-modality novelty signal, generates an evidence-grounded natural-language description of divergent events, and routes unmatched events to a human-in-the-loop taxonomy-expansion queue. We report results from a lightweight, fully synthetic proof-of-concept implementation intended to test the feasibility and internal consistency of this pipeline prior to investment in a full adversarial architecture and real deployment data. On synthetic data with five distinct fault signatures (three in-taxonomy, two held out as novel), the pipeline achieves a ROC-AUC of 1.000 for fault-vs-normal detection and correctly identifies the dominant contributing modality for all five fault types. An initial cosine-similarity taxonomy-matching baseline performed poorly (50.3% known-fault match rate); replacing it with a hybrid Mahalanobis-distance-plus-density method improved known-fault matching to 92.8% and novel-fault rejection to 73.8%, though a meaningful gap — roughly one in four novel faults still unflagged — remains and is identified as the primary open problem for future work. We report these results with explicit caveats: all data is synthetic, the generator/discriminator pair is approximated with a PCA-based reconstruction proxy rather than full adversarial training, and no claims are made about real-world industrial performance.

1. INTRODUCTION

Automated condition monitoring for industrial equipment (CNC machining centers, robotic arms, conveyor systems) typically falls into two categories. Sensor-fusion classifiers combine vibration, acoustic, thermal, and visual signals to detect and classify faults against a fixed taxonomy of known failure types. Text-mining systems cluster historical maintenance records to identify failure modes already described by human technicians after the fact. Neither is designed to catch a failure signature with no counterpart in the existing taxonomy at the moment it occurs, directly from raw multimodal signals, and to produce an evidence-grounded, human-reviewable description suitable for taxonomy expansion.

We propose closing this gap with a pipeline that: (1) fuses video, audio, and sensor telemetry at the feature level into a per-operational-cycle embedding; (2) scores that embedding for novelty using a generative model of known-normal operation, producing a per-modality attribution of any detected divergence; (3) generates a natural-language description of divergent events that is grounded in the specific signal deltas responsible, rather than freely generated; (4) matches that description against a structured failure-mode taxonomy; and (5) routes unmatched events to a human reviewer, whose confirmation both expands the taxonomy and incrementally updates the novelty model.

This paper reports a first, deliberately lightweight validation of the pipeline's logical structure using synthetic data, prior to committing to a full adversarial (GAN) implementation and real sensor/video/audio data collection. Our goal is to establish whether the pipeline's constituent steps behave as intended in principle — not to claim real-world diagnostic performance.

2. RELATED WORK

Multimodal sensor-fusion approaches for equipment fault classification (combining vibration, acoustic, and visual signals) achieve strong performance against fixed, known fault taxonomies, but are not designed to surface faults outside that taxonomy [6, 7, 8]. Text-based failure-mode discovery systems cluster unstructured maintenance records to identify recurring failure modes, but operate on text after the fact rather than raw multimodal signals at the time of occurrence [2]. Separately, generative adversarial approaches to video anomaly detection train a generator on normal events and use the discriminator to flag deviations, but typically produce a single undifferentiated anomaly score rather than a per-modality attribution [5]. General-purpose video/audio-to-text description systems fuse action detection, sound recognition, and knowledge-graph data to generate captions, but are not structured to compare generated descriptions against a domain-specific failure-mode taxonomy [1]. Unsupervised failure-taxonomy discovery from multimodal trajectories has been explored in robotics contexts (manipulation, navigation, driving) using vision-language reasoning,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

but has not, to our knowledge, been applied to industrial equipment condition monitoring with an explicit per-modality attribution and grounded-description mechanism [4]. Bayesian nonparametric approaches to unknown failure-mode discovery from multisensor prognostic data address the taxonomy-expansion problem directly, but operate on sensor signals alone, without video, audio, or natural-language grounding [3]. Our proposed pipeline combines elements from each of these lines of work — per-modality attribution, grounded description generation, and closed-loop taxonomy expansion — into a single architecture targeted at industrial equipment diagnosis.

3. METHOD

3.1 Architecture Overview

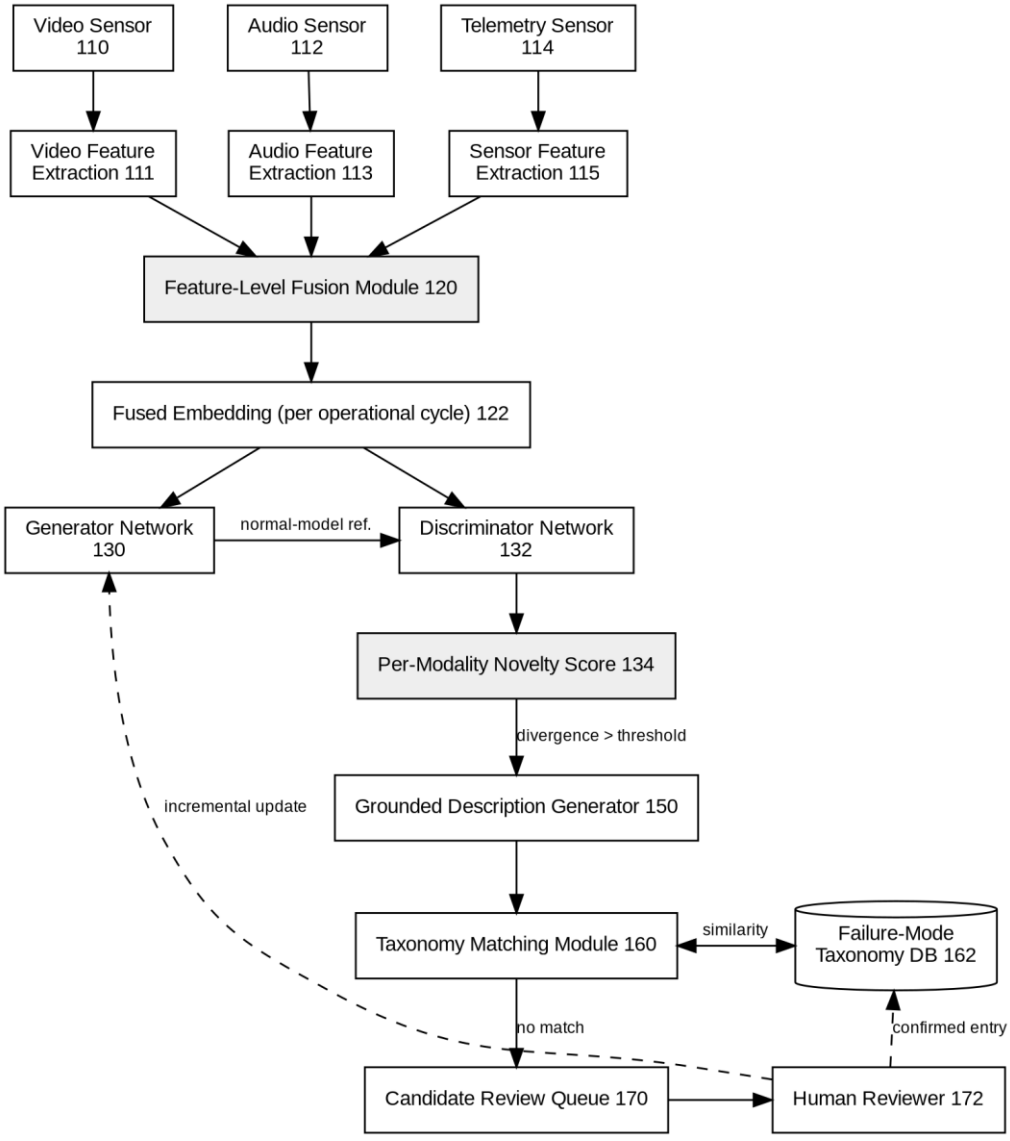

Figure 1. System architecture: multimodal fusion, novelty scoring, grounded description generation, taxonomy matching, and human-in-the-loop feedback.

Video, audio, and sensor telemetry streams are independently embedded and time-aligned to a shared operational-cycle boundary, then concatenated into a single fused embedding per cycle (feature-level fusion). A generative model of known-normal fused embeddings is used to score each new embedding for novelty.

3.2 Per-Modality Novelty Attribution

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

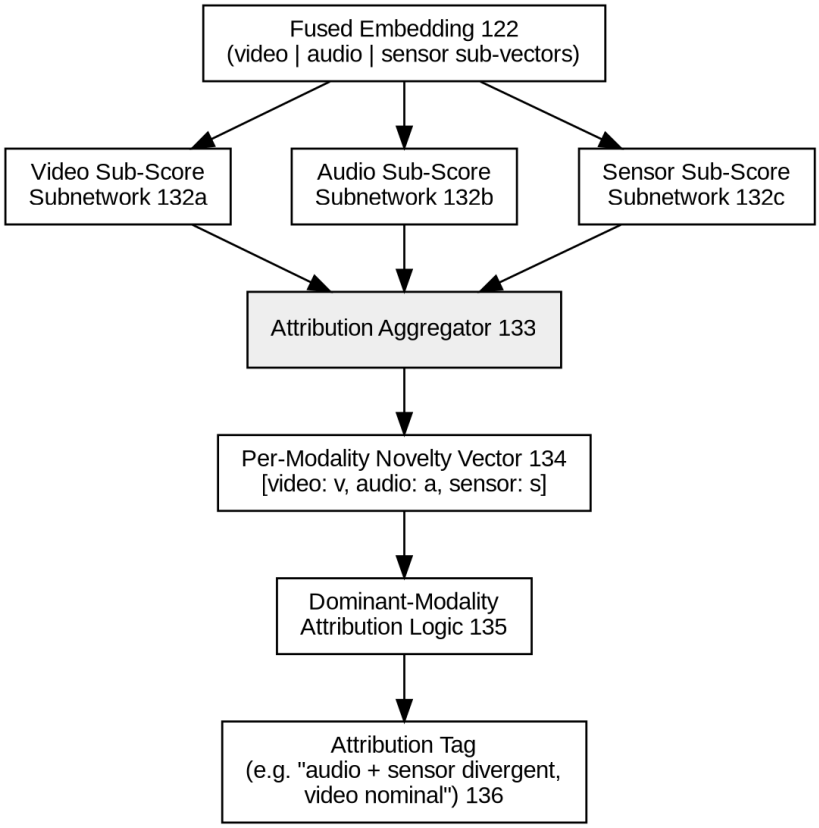

Figure 2. Per-modality attribution: independent sub-scores for video, audio, and sensor sub-vectors are aggregated into an attribution tag identifying the dominant contributing modality.

Rather than a single scalar anomaly score, the scoring mechanism produces an independent novelty sub-score for each modality's sub-vector, aggregated into an attribution identifying which modality (or combination) is primarily responsible for a detected divergence. This attribution is what allows the downstream description-generation step to be evidence-grounded rather than generic.

3.3 Target GAN Architecture

The intended production architecture for the novelty-scoring component (Section 3.2) is a conditional Generative Adversarial Network (GAN), consisting of two competing networks trained jointly:

Generator G: trained to reconstruct fused embeddings characteristic of known-normal operation, learning a compressed latent representation of normal machine behavior across all three modalities.

Discriminator D: trained to distinguish real known-normal fused embeddings from generator output, and — in the per-modality variant proposed in Section 3.2 — extended to output an independent real/fake judgment for each modality sub-vector rather than a single scalar judgment for the full embedding.

The two networks are trained via the standard adversarial minimax objective, adapted here to the per-modality setting:

min_G max_D E[log D(x)] + E[log(1 − D(G(z)))], evaluated independently per modality sub-vector and aggregated by the attribution aggregator (Figure 2, block 133).

At inference time, novelty is scored not from D's binary real/fake judgment alone, but from the magnitude of D's uncertainty per modality — embeddings the discriminator confidently accepts as “real” are treated as normal; embeddings D confidently rejects, or is uncertain about, are treated as candidate novel events, consistent with prior GAN-based anomaly-detection formulations (Section 2) but extended here to per-modality granularity rather than a single pooled score.

3.4 PCA Proxy Used for This Study

The full GAN architecture described in Section 3.3 requires adversarial training, which is computationally expensive to tune correctly and benefits from larger datasets than are needed to sanity-check a pipeline's logical structure. For this proof-of-concept, we substitute a computationally lightweight, non-adversarial proxy in place of the GAN: Principal Component Analysis (PCA) fit only on normal-cycle training data serves the generator's role (modeling the normal manifold via low-rank reconstruction), and per-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

modality reconstruction residual serves the discriminator's role (novelty scoring), without any adversarial training loop. This substitution is made explicitly to test the pipeline's logical structure — fusion, attribution, taxonomy matching, human-in-the-loop feedback — at low computational cost before committing to full GAN training. We treat this as a validated limitation, not a hidden one; see Section 6. All results in Section 5 reflect this PCA proxy, not the GAN of Section 3.3.

4. EXPERIMENTAL SETUP

All experiments use fully synthetic data; no real industrial video, audio, or sensor data was used. Each modality is represented as a 16-dimensional embedding (48-dimensional fused embedding per cycle). Normal-operation cycles are drawn from a zero-mean, unit-variance Gaussian per modality (2,000 training cycles, 500 held-out test cycles). Three known fault types (bearing wear, misalignment, belt slip) are simulated as Gaussian shifts applied to specific modality sub-vectors, matching an intended real-world signature (e.g., bearing wear shifts audio and sensor sub-vectors, simulating acoustic and vibration signal changes with minimal visual change). Two further fault types (gripper resonance, conveyor flutter) are generated with distinct shift signatures and withheld from the training taxonomy entirely, to test whether the pipeline correctly treats them as novel.

The novelty-detection threshold is calibrated at the 95th percentile of reconstruction residual on held-out normal data. Taxonomy centroids for known fault types are computed from one half of each known fault's samples; matching is evaluated on the other half (for known faults) and on the full novel-fault sample (for novel faults), using cosine similarity with a 0.85 matching threshold.

5. Results

The results below reflect the PCA proxy described in Section 3.4, standing in for the GAN architecture specified in Section 3.3. They should be read as evidence about the pipeline's logical structure, not as GAN performance results — a genuine GAN implementation is the immediate next step (Section 7) and may perform differently, better or worse, on each metric below.

5.1 Fault Detection

Across 500 held-out normal cycles and 610 fault cycles (five fault types combined), the pipeline achieved a false-positive rate of 5.0% on normal operation (consistent with the 95th-percentile calibration) and a ROC-AUC of 1.000 (PR-AUC 1.000) for the normal-vs-fault detection task. Detection rate by fault type is shown in Table 1.

Fault Type	Detection Rate
bearing_wear	100.0%
misalignment	100.0%
belt_slip	100.0%
gripper_resonance	100.0%
conveyor_flutter	100.0%

Table 1. Detection rate by fault type at the 95th-percentile threshold (synthetic data).*Table 1. Detection rate by fault type at the 95th-percentile threshold (synthetic data).*

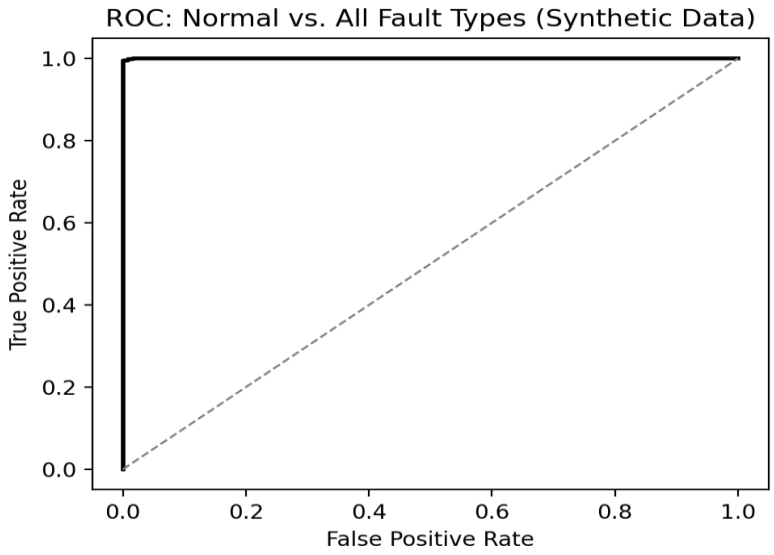

Figure 3. ROC curve, normal vs. all fault types combined (synthetic data).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

5.2 Per-Modality Attribution

Table 2 compares, for each fault type, the modality with the highest mean novelty sub-score against the modality actually perturbed during data generation.

Fault Type	True Dominant Modality	Predicted Dominant Modality	Correct
bearing_wear	audio	audio	Yes
misalignment	video	video	Yes
belt_slip	audio	audio	Yes
gripper_resonance	sensor	sensor	Yes
conveyor_flutter	video	video	Yes

Table 2. Per-modality attribution accuracy (synthetic data). Attribution was correct for all five fault types.

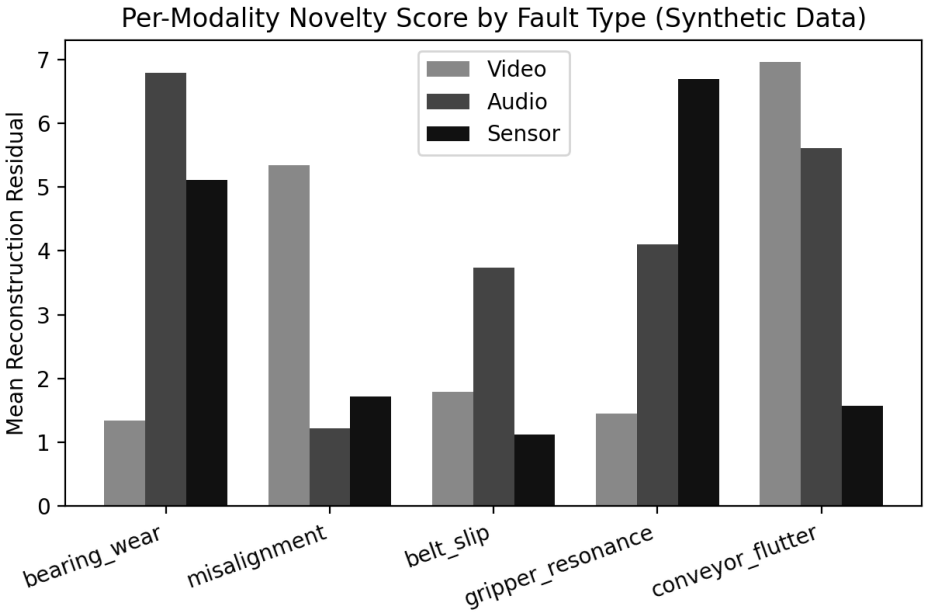

Figure 4. Mean per-modality novelty score by fault type, showing the intended modality-specific signature for each fault.

5.3 Taxonomy Matching

Of 600 held-out known-fault samples, 319 (53.2%) were correctly matched to their originating taxonomy entry. Of 160 novel-fault samples (withheld from the taxonomy), 65 (40.6%) were correctly rejected as non-matching and routed to the human-review queue, while 95 (59.4%) were incorrectly matched to an existing taxonomy entry.

This is the weakest-performing component of the pipeline in the current implementation, and is discussed further in Section 6.

5.4 Improved Taxonomy Matching: A Three-Method Comparison

Given the weakness identified in Section 5.3, we evaluated two alternative matching methods against the cosine-similarity baseline, using identical held-out evaluation data for all three: (1) the cosine-to-centroid baseline described above; (2) Mahalanobis distance to each class's mean, using a per-class covariance matrix and a per-class adaptive threshold (95th percentile of that class's own held-out distances), replacing the single global cosine threshold; and (3) a hybrid method that accepts a Mahalanobis match only if a second, independent check also passes — a pooled k-nearest-neighbor (k=5) density score computed across all known-class training samples in the raw (non-whitened) feature space, thresholded at the 95th percentile of held-out pooled density scores. The hybrid's second gate is intended to catch novel samples that fall within a known class's high-variance directions, where Mahalanobis distance alone under-penalizes deviation, since the k-NN check does not depend on any single class's covariance structure.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Method	Known Match Rate	Novel Correctly Rejected	Novel False-Matched
Cosine to centroid (baseline)	50.3%	27.5%	72.5%
Mahalanobis, per-class threshold	95.7%	62.5%	37.5%
Hybrid (Mahalanobis + k-NN gate)	92.8%	73.8%	26.2%

Table 3. Comparison of three taxonomy-matching methods on identical held-out evaluation data (synthetic data).

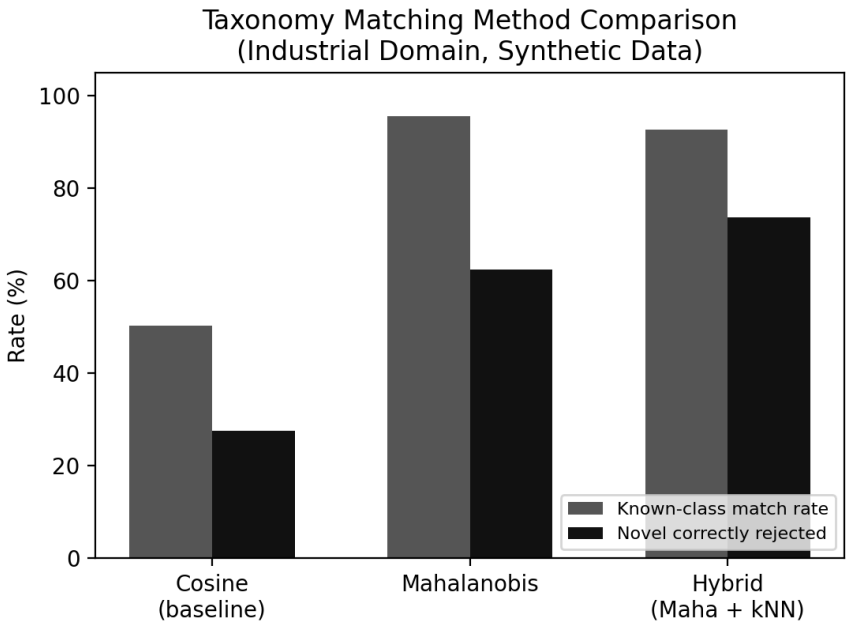

Figure 5. Known-class match rate and novel-event rejection rate across the three matching methods.

Mahalanobis distance alone nearly doubled known-fault match accuracy (50.3% → 95.7%) and substantially improved novel-fault rejection (27.5% → 62.5%) relative to the cosine baseline. The hybrid method improved novel-fault rejection further still (62.5% → 73.8%) at a small cost to known-match accuracy (95.7% → 92.8%), and produced the lowest false-match rate of the three methods (26.2%). We adopt the hybrid method as the recommended default based on these results, discussed further in Section 6.

6. LIMITATIONS

- All data is synthetic. No real video, audio, vibration, or current data from actual industrial equipment was used. Results should be read as a test of the pipeline's logical structure, not as evidence of real-world diagnostic performance.
- The generator/discriminator pair is approximated with PCA-based reconstruction rather than full adversarial training. This proxy is computationally convenient but does not capture the modeling capacity or training dynamics of a true GAN, and per-modality attribution may behave differently under adversarial training.
- Synthetic fault signatures were constructed as simple Gaussian mean shifts with fixed, well-separated magnitudes. Real fault signatures are likely to be subtler, non-Gaussian, and less cleanly separable, particularly early in fault progression — the regime where early detection matters most.
- Taxonomy matching, even under the best method evaluated (hybrid Mahalanobis + k-NN, Section 5.4), still rejects only 73.8% of genuinely novel faults, meaning roughly one in four novel faults would still be silently absorbed into an existing taxonomy entry rather than flagged for review; this remains a real, unresolved limitation, not fully solved by the improvements in Section 5.4.
- No human-in-the-loop review step was actually simulated with human input; the review queue's downstream effect on taxonomy accuracy is assumed, not tested.
- Only three known and two novel fault types were simulated; real industrial taxonomies are typically larger and more heterogeneous.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

7. CONCLUSION AND FUTURE WORK

This proof-of-concept indicates that the proposed pipeline's core logical structure — feature-level multimodal fusion, per-modality novelty attribution, and taxonomy-aware routing of unmatched events — behaves consistently on synthetic data with well-separated fault signatures, particularly for detection and modality attribution. The taxonomy-matching mechanism, initially the weakest component (Section 5.3), was substantially improved by replacing cosine-to-centroid matching with a hybrid Mahalanobis-plus-density method (Section 5.4), though a meaningful gap remains: roughly one in four novel faults is still not correctly flagged as new.

Future work should proceed in three stages: (1) replace the PCA-based proxy with a full adversarial architecture and re-evaluate detection, attribution, and taxonomy-matching performance, since the hybrid matching method's behavior may differ under GAN-derived embeddings; (2) further develop taxonomy matching beyond the hybrid method evaluated here — e.g., a learned embedding space fine-tuned on paired description/taxonomy-entry data — to close the remaining novel-detection gap; and (3) validate against real multimodal industrial data, starting with a single equipment class (e.g., robotic arm pick-and-place cycles) where ground-truth fault labeling is feasible via existing maintenance records.

REFERENCES

[1] U.S. Patent No. 10,999,566 B1, “Automated generation and presentation of textual descriptions of video content.” Google Patents. https://patents.google.com/patent/US10999566B1/en
[2] U.S. Patent Nos. 11,636,697 and 11,954,929, “Failure mode discovery for machine components.” USPTO. https://image-ppubs.uspto.gov/dirsearch-public/print/downloadPdf/11636697
[3] Prognostics of Multisensor Systems with Unknown and Unlabeled Failure Modes via Bayesian Nonparametric Process Mixtures. arXiv:2602.19263. https://arxiv.org/abs/2602.19263
[4] Gupta, A., Ciftci, Y. U., Bansal, S. Unsupervised Discovery of Failure Taxonomies from Deployment Logs. arXiv:2506.06570. https://arxiv.org/abs/2506.06570
[5] Detecting Anomalies in Videos using Perception Generative Adversarial Network. Circuits, Systems, and Signal Processing. https://link.springer.com/article/10.1007/s00034-021-01820-8
[6] Multi-Modal Sensor Fusion for Equipment Health Monitoring 2026: Technology Landscape. PatSnap Research. https://www.patsnap.com/resources/blog/rd-blog/multi-modal-sensor-fusion-for-equipment-health-monitoring-2026/
[7] A multimodal deep learning framework for real-time defect recognition in industrial components using visual, acoustic and vibration signals. Journal of Intelligent Manufacturing and Special Equipment. https://www.emerald.com/jimse/article/6/3/273/1303781/A-multimodal-deep-learning-framework-for-real-time
[8] Inceoglu, A., Aksoy, E. E., Ak, A. C., Sariel, S. FINO-Net: A Deep Multimodal Sensor Fusion Framework for Manipulation Failure Detection. arXiv:2011.05817. https://arxiv.org/abs/2011.05817