The Challenges of Multimodal IoT Data Fusion for Process Understanding and Monitoring

Zahra Ahmadi[1,*], Sönke Knoch[2,†], Massimiliano Sampaolo[3,†] and Mohsen Shirali[4,†]

[1]Research Centre for Information Systems Engineering (LIRIS), KU Leuven, Warmoesberg 26, 1000 Brussels, Belgium
[2]German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus, Saarbruecken, Germany
[3]School of Science and Technology, University of Camerino, Via madonna delle carceri, Camerino, Italy
[4]Computing Science and Engineering Department (INGI), UCLouvain, Louvain-la-Neuve, Belgium

Abstract
Multimodal IoT data fusion offers a promising way to improve process understanding and monitoring by combining heterogeneous IoT data across sensor, visual, acoustic, spatial, and contextual modalities. However, the integration of heterogeneous sources introduces several challenges, such as differences in data formats, sampling rates, and levels of abstraction. Despite the growing availability of multimodal IoT infrastructures, a unified framework that systematically addresses these challenges remains largely absent in the literature. Therefore, this paper focuses on the key challenges of multimodal IoT sources for process understanding and monitoring. It identifies four key sub-challenges: heterogeneous modality integration, fusion strategy selection, real-time activity and anomaly inference, and the construction of holistic and explainable process views. These sub-challenges are examined by reviewing research across various domains. Based on this analysis, the paper identifies key research gaps and opportunities and proposes a forward-looking approach for developing reliable and explainable multimodal IoT process intelligence systems.

Keywords
Multimodal IoT, Data fusion, Process understanding, Event abstraction, IoT challenges

1. Introduction

Fusion of multimodal Internet of Things (IoT) data for process understanding and monitoring is a technological approach that combines data from multiple IoT devices to gain comprehensive insights into industrial and operational processes [1]. This method integrates heterogeneous and multimodal IoT data, including sensor, event, spatiotemporal, visual, acoustic, and contextual data to offer complementary information on activities or cases, richer contexts, and greater awareness of the process [2]. Such complementary information has advantages across different domains. For example, manufacturing processes benefit from multimodal fusion through improved quality monitoring, where multiple sensor streams work together to detect faults and process deviations validated by using two chemical processes (Tennessee-Eastman and vinyl acetate monomer) [3]. Therefore, the authors propose a feature representation and fusion method, in which process time series signals are converted into image sequences. Features are represented from each senor's image sequence. A convolutional neural network is employed to fuse these features and generate new feature embeddings. The fault diagnose model is trained to optimize and balance the weight loss across the channels. In real-time process monitoring of complex environments such as transmission lines, multimodality can be used to combine data, such as temperature, vibration, and current. A fusion algorithm with feature extraction significantly reduces false alarms and improves fault-warning accuracy based on simulation experiments [1]. Beyond industrial settings, multimodal IoT and data fusion has shown strong potential in healthcare [4], smart

*Corresponding author.
†These authors contributed equally.
zahra.ahmadi@kuleuven.be (Z. Ahmadi); soenke.knoch@dfki.de (S. Knoch); massimilian.sampaolo@unicam.it (M. Sampaolo); mohsen.shirali@uclouvain.be (M. Shirali)
0000-0002-2313-5392 (Z. Ahmadi); 0000-0002-9097-8533 (S. Knoch); 0009-0003-2910-7979 (M. Sampaolo); 0000-0003-0584-4360 (M. Shirali)

environments [5], and logistics [6], where jointly interpreting heterogeneous data streams is essential for reliable process understanding.

However, integrating data and conducting analysis on it in real-world applications is challenging, as data is frequently collected from heterogeneous sources with diverse formats and structures. On the other side, involving more data leads to enhanced process logs which provide a more accurate data basis for process mining and further process analytics. Therefore, multimodal IoT data must be aligned, interpreted, and related to the activities, events, and cases that constitute the process. Without effective strategies for data fusion, generating deep insights and identifying deviations in process behavior is challenging: differences in data formats, sampling rates, semantic meaning, noise levels, and levels of abstraction must be handled [7]. These challenges become especially important when low-level sensor observations must be transformed into process-level knowledge in a meaningful and scalable way: (1) identifying which activity is being performed, (2) when an event starts or ends, (3) which case it belongs to, and (4) whether the observed behavior deviates from the expected process. Existing work often addresses these issues only partially, as sensor fusion, activity recognition, and process mining are commonly studied as separate problems. A unified and systematic treatment of multimodal IoT data fusion for process understanding is missing.

This gap motivates new algorithmic and methodological approaches, and has practical consequences, since poor integration appears differently for the stakeholders who depend on the resulting process insights: For process analysts, weak integration leads to incomplete, noisy, or semantically ambiguous event data, which directly affects the discovery, evaluation, and interpretation of process models [8]. For IoT system developers, the same problem appears as a technical challenge of synchronizing heterogeneous devices, managing missing or unreliable data, and maintaining robust data pipelines across diverse sensing infrastructures [9]. For industrial organizations, inadequate fusion can result in delayed or misleading process insights, reducing the effectiveness of quality control, predictive maintenance, compliance monitoring, and operational decision-making. For decision-makers, the issue is not only whether a model is accurate, but whether its outputs are trustworthy, explainable, and actionable, particularly in safety-critical domains such as healthcare, manufacturing, logistics, and smart environments. For researchers, the lack of unified integration frameworks limits the reproducibility, comparability, and transferability of methods across domains. Therefore, understanding the challenges of multimodal IoT data fusion is essential for developing reliable process representations that support meaningful visualization, behavioral analysis, deviation detection, and decision support.

The main contributions of this paper are therefore threefold. First, it frames multimodal IoT data fusion for process understanding and monitoring as a distinct research problem and systematically identifies four key sub-challenges: heterogeneous modality integration, fusion strategy selection, real-time activity and anomaly inference, and the construction of holistic and explainable process views. Second, it examines these sub-challenges through a cross-domain review of existing research in healthcare, human and natural behavior monitoring, smart environments, manufacturing, and logistics, thereby showing how similar integration problems appear across different application settings. Third, based on this analysis, the paper identifies key research gaps grouped into five categories and discusses corresponding opportunities for future work. To support the development of reliable and explainable multimodal IoT process intelligence systems, the paper further outlines forward-looking research directions organized as short-term actions, medium-term directions, and long-term visions.

The remainder of this paper is organized as follows. Section 2 introduces challenges related to multimodal IoT data fusion. In Section 3, representative states of the art, including fusion techniques and application areas, are reviewed. Section 4 discusses gaps and opportunities. Proposed approaches and visions are presented in Section 5. The paper ends with conclusion and future work in Section 6.

2. Sub-Challenges and Problem Decomposition

Multimodal IoT fusion can serve different goals, including activity extraction, context enrichment, anomaly detection, process monitoring, and decision support. Depending on the application setting,

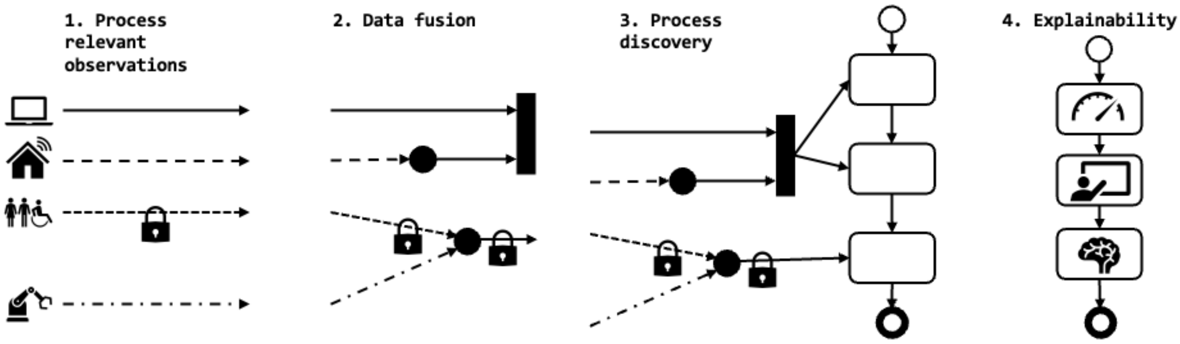

Figure 1: The four sub-challenges reflecting multimodal IoT data fusion from data source to process level.

modalities may contribute differently: some primarily support activity inference, while others provide contextual or complementary information. This variation shows that multimodal fusion is not a single uniform task, but is shaped by the monitoring objective, the available sensing infrastructure, and the domain context. Across these settings, however, the same general challenge remains: heterogeneous IoT data streams must be transformed into meaningful process-level information. This transformation decomposes into several interacting sub-challenges, where decisions made at one stage influence the quality and usefulness of the next. The following paragraphs discuss these sub-challenges in detail, which are presented in Figure 1.

SC1-Integrate diverse IoT modalities to support comprehensive process understanding: Multimodal IoT systems bring together visual, acoustic, textual, and sensor-based devices to collect data and describe processes from different perspectives. The main difficulty lies in synchronizing and semantically aligning heterogeneous observations that differ in format, sampling rate, temporal granularity, and meaning. For instance, wearable devices may continuously capture physiological or motion signals, while smartphones and ambient sensors provide more sparse or event-driven observations. To effectively use these observations for process understanding, it is essential to interconnect these observations to the same activities, events, or cases, even when some data inputs are incomplete, asynchronous, or partially conflicting [10]. Furthermore, this integration faces additional constraints due to privacy regulations, fragmented systems, competing interests among different parties involved, and limited data sharing, particularly in sensitive fields like healthcare [4]. This represents a foundational sub-challenge: if the various modalities are poorly synchronized or semantically misaligned and misinterpreted, subsequent steps in data integration may produce unreliable activity labels, incorrect anomaly detection results, or misleading process models.

SC2-Employ fusion strategies, multimodal deep learning, and hybrid reasoning techniques to combine heterogeneous data sources: After modalities get aligned and connected to process-relevant observations, the next challenge is deciding how they should be fused. This choice is important because modalities do not contribute equally: some may be highly predictive for a given activity or event, while others mainly provide contextual or confirmatory evidence. Early fusion combines raw data or low-level features from multiple modalities before model training or inference. This strategy can capture fine-grained cross-modal relationships, but it usually requires strongly-synchronized data streams and consistent feature quality. In contrast, late fusion first analyzes each modality separately, using modality-specific models to produce outputs such as activity labels, machine states, risk scores, or anomaly indicators. These outputs are then merged into a final decision [11]. Late fusion is more flexible when modalities are asynchronous or unavailable, but it may overlook cross-modal interactions that are only visible when low-level signals are considered together. Hybrid fusion combines both principles by integrating low-level representations with higher-level decisions, and can be supported by multimodal deep learning, attention mechanisms, shared embedding spaces, and reasoning-based techniques [12]. Such strategies are useful for balancing complementary evidence, handling unequal modality importance, and supporting robust inference. Fusion strategy selection directly affects robustness, explainability, and real-time inference capabilities.

SC3-Infer activities, contexts, and anomalies in real time, enabling proactive decision-making and adaptive process monitoring: Once multimodal observations have been integrated and fused, the resulting representations must be translated into process-relevant events, contexts, and anomalies. This is challenging because IoT streams are often continuous, noisy, and low-level, whereas process mining and monitoring require discrete, meaningful events, such as activity occurrences, contextual states, or case-level updates. In human behavior monitoring, for example, smartphones, wearables, and ambient sensors capture continuous movement, interactions, and physiological responses, but human behavior rarely appears as a sequence of clearly separated activities. Human activities often overlap, evolve gradually, and include concurrent micro-behaviors, making activity boundaries ambiguous [13]. Similar difficulties arise in industrial and logistics settings, where gradual changes in sensor signals may indicate emerging faults or process deviations before they become explicit events.

This sub-challenge depends directly on the previous two: reliable real-time inference requires well-integrated data streams and fusion strategies that can handle changing modality relevance. These inferred elements enable adaptive monitoring, early warning, decision support, and PM.

SC4-Provide a holistic and explainable view of IoT-driven business processes, capturing both the physical and digital dimensions of execution: Multimodal fusion should ultimately support explainable process analysis. This means that the outputs of fusion and inference must be organized into process representations that analysts can inspect, explain, and relate to the actual execution of a process. Process representations should integrate both digital traces and physical observations. However, the same observations may lead to different process interpretations depending on how activities, cases, contexts, and outcomes are defined. For example, in animal behavior monitoring, a case may correspond to an individual animal, a group, or a migration episode, leading to overlapping and multi-level process interpretations [14]. Therefore, holistic process views must be explainable and domain-valid, allowing analysts to understand not only what happened, but also why a particular interpretation was produced. This sub-challenge builds on the previous ones: explainable process views depend on reliable integration, fusion, and inference, while also providing feedback for validating earlier stages.

The four sub-challenges are closely connected: Poor modality integration can reduce the quality of fusion, unsuitable fusion strategies can limit the reliability of real-time activity and anomaly inference, and inaccurate inference can lead to incomplete or misleading process views. However, these dependencies are not strictly linear. Problems observed during activity inference may reveal weaknesses in earlier integration steps, while explainability requirements may influence the choice of fusion strategy, particularly in domains where decisions must be justified, e.g., healthcare, manufacturing, and logistics. Therefore, multimodal IoT data fusion for process understanding should be treated as an iterative process in which integration, fusion, inference, and explanation continuously inform each other.

3. State of the Art

3.1. Research Perspective

Academic research on multimodal IoT data fusion for process understanding and monitoring can be grouped into several main research directions: event abstraction, multimodal fusion strategies, process discovery, uncertainty handling, predictive monitoring, scalability, and explainability.

Event Abstraction from Sensor Data: The first research direction focuses on event abstraction, which serves as a bridge between low-level IoT data and PM. A central challenge in this research area is converting continuous, low-level device measurements and readings into discrete process events. The typical pipeline for event abstraction usually follows a sequence of steps: preprocessing, segmentation or clustering, labeling, and event-log construction. Solutions include pattern-based abstraction, clustering of sensor streams, and machine-learning-based activity recognition [15]. For example, IoT Miner provides an automated pipeline which includes clustering and labeling using Large Language Models (LLMs) to generate high-level logs from sensor data [16]. In addition, LLM-based event abstraction and integration for IoT-sourced event data is particularly useful when semantics are missing or fragmented [17].

Multimodal Data Fusion Strategies: After sensor data has been processed, another major research direction concerns how different modalities should be combined. Modern approaches combine heterogeneous modalities using three main paradigms: early or feature-level fusion, late or decision-level fusion, and hybrid fusion [13], which are often effective in activity recognition or behavior detection before PM. Surveys on multimodal event detection and fusion further support the idea that events can first be inferred from heterogeneous modalities and then used as input for PM [18]. Beyond these general paradigms, several model-based approaches have been developed to handle uncertainty, semantics, and temporal dependencies. Probabilistic models such as Dynamic Bayesian Networks explicitly model uncertainty and temporal dependencies. Semantic aggregation approaches map heterogeneous sensor outputs into ontology-based representations, such as RDF knowledge graphs. Tensor-based multimodal fusion methods encode spatiotemporal features and integrate modalities such as traffic cameras, geomagnetic sensors, and social media streams into low-rank representations.

Process Discovery from IoT Data: Another research direction focuses on transforming IoT data into process-oriented representations that can support process discovery and monitoring. Frameworks such as those proposed by Koschmider [19] and Mangler [9] introduce pipelines that include preprocessing, event extraction, case correlation, activity abstraction, and process discovery. These explicitly bridge the gap between physical sensor observations and business process models, which is important because multimodal IoT data does not automatically provide process knowledge; it must first be interpreted in relation to activities, cases, resources, and timestamps before PM can be applied. Streaming PM [20] extends this perspective to real-time environments, enabling continuous monitoring and adaptation as new IoT events are generated.

Handling Noise, Missing Data, and Uncertainty: State-of-the-art methods employ probabilistic models, imputation techniques, robust machine learning algorithms, and error-correction approaches for IoT-sourced event logs [21, 10]. Because IoT devices are often noisy, asynchronous, and intermittently available, uncertainty handling is essential before sensor data can be converted into reliable process events. This is especially important and studied in healthcare, manufacturing, smart environments, and logistics, where incorrect event abstraction can lead to misleading process models or false alarms. Understanding the types and sources of errors, and adapting correction methods accordingly, can improve the reliability of event logs extracted from raw sensor readings [10]. In this sense, uncertainty handling supports both reliable event abstraction and trustworthy process monitoring.

Multimodal Predictive Process Monitoring: In addition to discovering and monitoring current processes, research also studies how IoT data can improve prediction [22]. Multimodal predictive process monitoring enriches traces with text, audio, video, and sensor features for prediction and monitoring. Reviews specifically on integrating multimodal data into predictive process monitoring focus on combining logs with language and vision data [23].

Scalability and Real-Time Processing: Edge computing, Federated Learning (FL) and distributed analytics are widely proposed to handle massive data streams and enable decentralized analytics and privacy-preserving multimodal learning across distributed infrastructures. Systems such as EdgeMiner perform partial processing near data sources, reducing latency and communication overhead while enabling real-time decision support [24]. The ongoing studies in this direction are particularly relevant when multimodal data is distributed across devices, organizations, or locations.

Explainability and Context Integration: Research increasingly emphasizes that multimodal process monitoring must be explainable and context-aware, as its output should be readable for clinicians, operators, analysts, and decision-makers. Hybrid approaches combining data-driven models with domain knowledge are used to produce interpretable and domain-valid results [25]. Context-aware PM incorporates environmental, temporal, organizational, and situational factors to improve the validity of discovered models and the interpretation of process behavior. Therefore, recent research increasingly combines machine learning with semantic reasoning, process models, ontologies, and expert knowledge to make monitoring results more understandable, transparent, and actionable.

3.2. Industry and Practice Perspective

Implemented real-world systems that rely on multimodal IoT fusion can be broadly categorized according to the nature of the processes they monitor into the following three groups:

Human-centric processes: Human-centric processes are organized around people as the main process participants or cases, such as patients, workers, or residents. In these settings, multimodal IoT fusion is primarily used to enhance monitoring, guidance, and decision-making. In healthcare, one of the most mature application areas, heterogeneous sources such as Electronic Health Records, wearables, medical imaging, environmental sensors, and clinical observations are combined to provide a more complete view of patient status and care delivery [4]. For example, multimodal PM has been used in Surgery AI to support real-time surgical conformance checking and guidance [26]. In human behavior monitoring using smart homes and ambient assisted living systems multimodality is utilized for improved recognition of daily activities and behavioral routines. Similarly, digital trace data from mHealth applications and wearable sensors are used in Human Behavior Mining (HBM) framework to capture and analyze dynamic patterns of human behavior over time [27]. In these systems, the practical value of multimodal fusion lies in linking fragmented observations of human behavior to meaningful activities, deviations, and decision-support outputs [10, 17].

Natural processes: Natural processes refer to biological, ecological, or environmental phenomena whose execution is not directly controlled by an information system but can be observed through sensing technologies. In practice, multimodal IoT fusion is used to make such processes observable by combining complementary sensing streams. For example, animal behavior monitoring often integrates accelerometry, positional, acoustic signals, environmental conditions, and sometimes visual observations to classify behaviors [28]. These classified behaviors can serve as event abstractions for later process discovery or behavioral trajectory analysis in veterinary and livestock monitoring frameworks.

Cyber-Physical Systems (CPS): CPSs encompass industrial, urban, logistics, and infrastructure processes in which physical operations are closely integrated with digital monitoring, control, and decision-support systems. Across these categories, multimodal IoT fusion is used in practice to improve observability, support real-time monitoring, and connect physical-world behavior to process-level representations. This category includes manufacturing systems, smart cities, transportation networks, logistics operations, and supply chains. In such environments, multimodal IoT fusion is used to monitor the state of machines, products, vehicles, infrastructure, and environmental conditions in real time.

In *smart cities*, for instance, traffic cameras, geomagnetic or loop sensors, GPS traces, environmental monitoring stations, and social media streams are combined to support traffic flow estimation, congestion prediction, incident detection, and urban decision-making. These modalities differ in scale and granularity, but together they provide a richer view of urban processes.

In *manufacturing*, multimodal fusion supports activity recognition, quality monitoring, process automation, and conformance checking. Rinderle-Ma and Mangler highlight the complementarity between process automation and PM in digital manufacturing environments [29]. Roitberg et al. present a multimodal vision-based interface for recognizing human activities in industrial manufacturing processes [30], while Knoch et al. propose extracting traces from video data to support process discovery and conformance checking in manual assembly [31]. These examples show how physical execution data, especially from cameras and sensors, can be transformed into process traces.

In *logistics* and supply-chain contexts, IoT technologies are used to monitor goods, vehicles, storage conditions, and transportation workflows. Temperature-controlled transportation systems, for example, use sensors embedded in smart packaging together with BPMN-based workflows to monitor specimen conditions during transport [32]. Machine learning-based decision-support systems further analyze logistics sensor data for tasks such as route optimization, demand prediction, and inventory management [33]. More generally, IoT-enhanced BPM frameworks combine semantic ontologies, microservices, and connected infrastructures to integrate IoT data with business processes [34]. In practice, these systems demonstrate how multimodal IoT fusion can connect physical execution with digital process models, enabling monitoring, optimization, and decision support across distributed operations.

3.3. Synthesis and Comparison

The research and practice perspectives show both strong alignments and important mismatches. Academic research mainly focuses on methodological challenges, such as event abstraction, multimodal representation learning, activity recognition, uncertainty handling, process discovery, predictive monitoring, and explainability. Practical implementations, in contrast, are often driven by operational goals such as real-time monitoring, fault detection, predictive maintenance, patient safety, logistics traceability, activity tracking, and decision support. Both perspectives acknowledge that single data sources are often insufficient for understanding complex processes and that combining sensor, contextual, visual, acoustic, textual, and system-generated data can improve monitoring accuracy and robustness.

However, a major mismatch appears in the assumptions made about data availability and quality. Research studies often rely on datasets that are synchronized, labeled, cleaned, and accessible for experimentation. In practice, multimodal IoT environments frequently involve missing data, unreliable sensors, fragmented infrastructures, proprietary systems, privacy constraints, and limited ground truth. As a result, methods that perform well in controlled research settings may be difficult to transfer directly into operational environments. Another important mismatch concerns explainability and trust. Many research approaches use deep learning and complex multimodal models to improve predictive performance, but practitioners often require transparent explanations for alerts, anomalies, and recommendations. This is especially important in healthcare, manufacturing, logistics, and other safety-critical domains, where incorrect or unexplained decisions can have significant consequences. Therefore, explainability is not only a methodological research objective but also a practical requirement for adoption. Overall, Research provides advanced methods for individual parts of the problem, while practice demonstrates the need for robust, scalable, and trustworthy solutions that work under real-world constraints. This indicates the need for future work that bridges methodological advances with deployable process intelligence systems.

4. Research Gaps and Opportunities

The analysis of research and industry perspectives shows that multimodal IoT data fusion has strong potential for process understanding and monitoring, but several limitations and gaps still prevent its systematic use. These limitations do not concern only the performance of individual fusion models, but also concern how raw data streams are transformed into process-level data, how reliable and timely these transformations are, and how the resulting insights can be explained, evaluated, and transferred across domains. Accordingly, the main gaps can be summarized in three interrelated areas:

Gap 1: From raw multimodal streams to structured process-level data. A central gap concerns the transformation of raw multimodal IoT streams into structured event data that can support typical PM goals: process discovery, conformance checking, and monitoring. Existing approaches often focus on a specific domain, sensor setup, or activity recognition task, which makes them difficult to reuse in other settings. For instance, methods developed for human behavior monitoring, smart environments, industrial assembly, or logistics typically define their own preprocessing, segmentation, labeling, and case-correlation procedures. As a result, there is still no general framework that explains how multimodal observations should be abstracted into process events across domains.

This gap exists because multimodal IoT data is not naturally process-oriented. IoT devices capture signals, states, movements, images, sounds, or interactions, while PM requires event logs with activities, timestamps, cases, resources, and context. Moving from one representation to the other requires several design decisions: which signal changes correspond to an event, how event boundaries are defined, how multiple sensor observations are combined, and how events are assigned to cases. If these decisions are made ad hoc, the resulting event logs may be difficult to compare, reproduce, or validate.

The implication is that multimodal fusion may produce useful predictions without necessarily producing reliable process knowledge. A model may correctly classify an activity, but if the activity cannot be linked to a case, timestamp, or process context, its value for process understanding remains limited. As an opportunity, future research should therefore develop reusable event abstraction frameworks that cover preprocessing, segmentation, feature extraction, semantic labeling, case correlation, and event-log construction. Such frameworks should also include evaluation criteria for fusion quality, event abstraction quality, process model fidelity, monitoring accuracy, and explainability. Emerging approaches such as LLM-based IoT event abstraction and IoT Miner provide promising directions for moving from low-level sensor data to higher-level event logs [17, 16]. However, these efforts still need to be connected into end-to-end pipelines that can be evaluated across domains.

Gap 2: Reliable and timely process monitoring under real-world IoT conditions. Gap 2 concerns the reliability and timeliness of multimodal IoT data when used for online monitoring. Many research methods are evaluated on datasets that are relatively clean, synchronized, and complete. In real deployments, IoT data is often noisy, missing, asynchronous, delayed, or affected by changing device configurations, which is especially critical in real-time process monitoring, where incorrect event abstraction or delayed anomaly detection can lead to misleading alerts and poor operational decisions.

This gap exists because most methods treat uncertainty, streaming, and process abstraction as separate problems. Imputation methods may address missing values, activity recognition models may infer labels, and PM techniques may analyze event logs, but fewer approaches combine them into a single uncertainty-aware monitoring pipeline. Real-time settings also introduce additional difficulties, e.g. concept drift, changing sensor reliability, limited ground truth, and resource constraints at the edge.

Current approaches may be accurate in offline experiments but fragile in operational environments. For safety-critical domains such as healthcare, manufacturing, logistics, and smart infrastructure, this fragility can reduce trust and limit adoption. The opportunity for future works to address this gap is to combine uncertainty-aware modeling, probabilistic alignment, sensor reliability estimation, robust imputation, and confidence-aware event abstraction. In addition, streaming abstraction and incremental process monitoring should be integrated with drift detection and edge-based analytics. Early examples, such as real-time surgical conformance checking, show the potential of online multimodal monitoring [26], but broader solutions are still needed for general multimodal IoT process intelligence.

Gap 3: Explainable, transferable, and evaluable multimodal process intelligence. A third gap concerns the ability to produce process insights that are explainable, transferable, and systematically evaluated. Much multimodal research focuses on classification, prediction, or anomaly detection, while PM emphasizes interpretability, behavioral analysis, and process improvement. The connection between these perspectives remains underdeveloped. In particular, multimodal fusion models often produce outputs such as labels or scores, but they do not always explain how these outputs relate to process behavior, deviations, resources, cases, or outcomes.

This gap exists partly because physical and digital process views are still difficult to integrate. IoT data captures what happens in the physical world, while traditional process analysis often relies on digital system logs. Bridging these views requires representations that connect sensor observations with business objects, activities, resources, contextual conditions, and process outcomes. It also requires explanation mechanisms that make fusion results understandable to analysts, operators, clinicians, and decision-makers. This is particularly important when highly accurate black-box models are used in domains where decisions must be justified.

The implication is that multimodal systems may support prediction without supporting understanding. This limits their usefulness for root-cause analysis, compliance checking, process redesign, and human decision-making. Future research should therefore focus on explainable multimodal PM, combining data-driven models with semantic reasoning, ontologies, domain knowledge, and process models. The goal should be not only to improve predictive accuracy, but also to explain why a behavior was detected, why something is considered anomalous, and how it affects the process.

This gap is also visible in evaluation and transferability. There is a lack of standard benchmarks and datasets for multimodal IoT process analysis. Existing datasets are often synthetic, domain-specific, or limited in the modalities they include [35]. Moreover, similar challenges appear across particular domains such as healthcare and logistics, but these communities often develop methods independently. Future work should define benchmark tasks for event abstraction quality, process fidelity, monitoring accuracy, robustness, explainability, and real-time performance. Comparative studies are also needed to identify which abstraction and fusion strategies can transfer across domains and which require domain-specific adaptation.

Together, these gaps indicate that multimodal IoT data fusion for process understanding should move beyond isolated models and the field requires integrated pipelines that transform heterogeneous data streams into reliable, explainable, and process-oriented representations. Addressing these gaps would support the development of multimodal IoT process intelligence systems that are not only accurate but also robust, interpretable, reusable, and suitable for real-world deployment.

5. Proposed Approach

5.1. Conceptual Vision

We envision a process-oriented multimodal IoT intelligence framework that transforms heterogeneous sensor observations into reliable, explainable, and actionable process insights. Rather than treating multimodal fusion only as a prediction task, the framework positions fusion as part of a broader pipeline for process understanding and monitoring. The framework follows five connected stages. First, multimodal data are acquired according to the monitoring objective and the process perspective, using sources such as sensors, cameras, wearables, acoustic devices, enterprise systems, and other IoT-enabled sources. Second, these streams are prepared and integrated by addressing synchronization, alignment, contextual enrichment, and data quality issues. Third, fusion and event abstraction techniques are applied to convert low-level observations into process-relevant events, activities and contexts. Fourth, the resulting event data are used for process discovery, conformance checking, predictive monitoring, and anomaly detection. Finally, the outputs are translated into explainable process views and decision-support information for analysts, operators, and decision-makers.

The key principle of this vision is that fusion should be process-oriented, not only prediction-oriented. It also requires understanding how physical observations relate to digital traces, how events belong to cases, how deviations emerge, and how results can be trusted by stakeholders. Therefore, uncertainty handling, domain knowledge, flexible event abstraction, and explainability should be embedded throughout the pipeline. The framework should also be iterative: monitoring results and explanations can reveal weaknesses in earlier stages and guide refinements in data integration, fusion strategy, or event abstraction. Based on this conceptual vision, the following short-term, medium-term, and long-term actions are proposed.

5.2. Short-Term Actions

Refer to improvements that can be implemented immediately with existing methods and datasets.

- **Action 1:** Establish standardized multimodal event abstraction pipelines that include preprocessing, segmentation, feature extraction, semantic labeling, and event log construction. These pipelines should compare early, late, hybrid, and reasoning-based fusion strategies under similar conditions to clarify their suitability for different data characteristics.
- **Action 2:** Improve robustness to noisy, missing, and asynchronous data by incorporating uncertainty-aware models, probabilistic alignment, and robust preprocessing strategies that can handle changing sensor configurations and intermittent data streams.
- **Action 3:** Include explainability as an evaluation criterion. Researchers should evaluate not only predictive accuracy, but also whether the resulting events, process models, anomalies, and recommendations are understandable to domain experts.
- **Action 4:** Use domain knowledge during event abstraction through rules, ontologies, expert feedback, or semantic constraints to reduce meaningless event labels and improve the link between sensor signals and process concepts.
- **Action 5:** Evaluate event abstraction before PM by checking whether the abstracted events are meaningful, correctly segmented, and semantically valid.

5.3. Medium-Term Directions

Medium-term directions require shared resources, reusable frameworks, and broader validation beyond isolated case studies.

- **Direction 1:** Develop benchmark datasets and evaluation frameworks for multimodal process analysis across various domains. These datasets should include multiple modalities, contextual information, uncertainty annotations, and domain-specific ground truth where possible.
- **Direction 2:** Conduct cross-domain transfer studies to examine whether event abstraction and fusion methods developed in one domain can be adapted to another.
- **Direction 3:** Create reusable representations for IoT-enhanced event logs that support multimodal context, changing sensors, uncertain events, and links between physical observations and digital process records. Such representations can support interoperability and move the field toward domain-agnostic standards.
- **Direction 4:** Integrate PM with digital twins to support real-time monitoring, simulation, prediction, and explanation in manufacturing, logistics, healthcare, and smart cities.

5.4. Long-Term Vision

Long-term visions require structural advances in infrastructure, governance, and intelligent system design.

- **Vision 1:** Enable real-time, explainable multimodal process intelligence systems that continuously adapt to new behaviors, changing environments, new devices, and process drift. This requires advances in streaming PM, adaptive learning and explainable models.
- **Vision 2:** Establish cross-organizational data-sharing infrastructures for multimodal IoT processes that span multiple actors. Secure and standardized infrastructures are needed to support data sharing, traceability, and process analysis across organizational boundaries.
- **Vision 3:** Unify symbolic, statistical, and deep learning approaches to combine interpretability, uncertainty handling, and representation learning in a single process-oriented framework.
- **Vision 4:** Develop governance frameworks for ethical and trustworthy multimodal monitoring by combining privacy-preserving analytics, edge analytics, and secure aggregation, with principles of consent, fairness, accountability, transparency, and responsible use.

6. Conclusions and Future Work

This paper examined challenges of multimodal IoT data fusion for process understanding and monitoring. It identified key sub-challenges, including heterogeneous data integration, fusion strategy selection, event abstraction, real-time activity and anomaly detection, and explainable process representation. It also showed that these challenges appear across several domains, such as healthcare, human behavior monitoring, natural process monitoring, smart environments, manufacturing, and logistics. The gaps and opportunities beside the state of the art are discussed, and proper actions and visions are proposed for these challenges. Future work should focus on developing standardized event abstraction pipelines, robust methods for uncertainty-aware analysis, benchmark development, reusable IoT-enhanced event-log representations, and explainable multimodal PM frameworks.

References

[1] Q. Kong, J. Wang, G. Zhang, B. Wang, X. Zhang, Research on multimodal mechanism data fusion technology in transmission lines, in: 2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering (ICEACE), IEEE, 2024, pp. 1104–1108.

[2] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: an overview of methods, challenges, and prospects, Proceedings of the IEEE 103 (2015) 1449–1477.

[3] Y. Shi, Z. Li, B. Yang, A multi-sensor feature representation and fusion method for data-driven industrial process fault diagnosis, IEEE Journal of Emerging and Selected Topics in Industrial Electronics (2025).

[4] T. Shaik, X. Tao, L. Li, H. Xie, J. D. Velásquez, A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom, Information Fusion 102 (2024) 102040.

[5] Y. Kirikkayis, F. Gallik, M. Winter, M. Reichert, Bpmne4iot: A framework for modeling, executing and monitoring iot-driven processes, Future Internet 15 (2023). URL: https://www.mdpi.com/1999-5903/15/3/90. doi:10.3390/fi15030090.

[6] D. Crnjac Milić, I. Dujmenović, M. Peko, An approach to the application of the internet of things in logistics, Tehnički glasnik 17 (2023) 134–140.

[7] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 423–443.

[8] Y. Bertrand, R. Van Belle, J. De Weerdt, E. Serral, Defining data quality issues in process mining with iot data, in: International Conference on Process Mining, Springer, 2022, pp. 422–434.

[9] J. Mangler, R. Seiger, J.-V. Benzin, J. Grüger, Y. Kirikkayis, F. Gallik, L. Malburg, M. Ehrendorfer, Y. Bertrand, M. Franceschetti, et al., From internet of things data to business processes: Challenges and a framework, in: Internet of Things Meets Business Process Management: A Synergistic Integration, Springer, 2026, pp. 25–61.

[10] M. Shirali, Z. Ahmadi, C. Fernández-Llatas, J. L. Bayo Montón, G. Di Federico, An interactive error-correcting approach for iot-sourced event logs, ACM Transactions on Internet of Things 5 (2024) 1–30.

[11] M. Pawłowski, A. Wróblewska, S. Sysko-Romańczuk, Effective techniques for multimodal data fusion: A comparative analysis, Sensors 23 (2023) 2381.

[12] M. McKinney, A. Garland, D. Cillessen, J. Adamczyk, D. Bolintineanu, M. Heiden, E. Fowler, B. L. Boyce, Unsupervised multimodal fusion of in-process sensor data for advanced manufacturing process monitoring, Journal of Manufacturing Systems 78 (2025) 271–282.

[13] J. Wu, Y. Feng, P. Sun, Sensor fusion for recognition of activities of daily living, Sensors 18 (2018) 4029.

[14] Z. Ahmadi, Y. Bertrand, M. I. Pozo Romero, E. Serral, Analysing the foraging behaviour of bees using process mining: A case study, in: International Conference on Process Mining, Springer, 2023, pp. 5–18.

[15] N. Ravi, N. Dandekar, P. Mysore, M. L. Littman, Activity recognition from accelerometer data, in: Aaai, volume 5, Pittsburgh, PA, 2005, pp. 1541–1546.

[16] E. Brzychczy, U. Jessen, K. Kluza, S. Sriram, M. V. Nettelnstroth, Iot miner: Intelligent extraction of event logs from sensor data for process mining, arXiv preprint arXiv:2509.05769 (2025).

[17] M. Shirali, M. F. Sani, Z. Ahmadi, E. Serral, Llm-based event abstraction and integration for iot-sourced logs, in: International Conference on Business Process Management, Springer, 2024, pp. 138–149.

[18] M. Mondal, M. Khayati, H.-Â. Sandlin, P. Cudré-Mauroux, A survey of multimodal event detection based on data fusion: M. mondal et al., The VLDB Journal 34 (2025) 9.

[19] A. Koschmider, D. Janssen, F. Mannhardt, Framework for process discovery from sensor data, in: 10th International Workshop on Enterprise Modeling and Information Systems Architectures, EMISA 2020, CEUR-WS. org, 2020, pp. 32–38.

[20] A. Burattin, Streaming Process Mining, Springer International Publishing, Cham, 2022, pp. 349–372. URL: https://doi.org/10.1007/978-3-031-08848-3_11. doi:10.1007/978-3-031-08848-3_11.

[21] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, O. Tabona, A survey on missing data in machine learning, Journal of Big data 8 (2021) 140.
[22] Y. Bertrand, A. Stevens, B. Deforce, J. De Smedt, J. De Weerdt, E. Serral, Approaches for iot-enhanced predictive process monitoring, Process Science 2 (2025) 7.
[23] A. Hafner, N. Kim, H. Wittges, From logs to language and vision: A review on integrating multimodal data into predictive process monitoring, in: 59th Hawaii International Conference on System Sciences (HICSS 59), University of Hawaii at Manoa, 2026, pp. 5641–5650.
[24] J. Andersen, P. Rathje, C. Imenkamp, A. Koschmider, O. Landsiedel, Edgeminer: Distributed process mining at the data sources, in: Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, 2025, pp. 705–713.
[25] Z. Ahmadi, M. Shirali, J. De Weerdt, E. Serral, Context-enriched process discovery from iot data sources for human behavioral monitoring, in: IFIP Working Conference on The Practice of Enterprise Modeling, Springer, 2025, pp. 209–226.
[26] A. Gavric, D. Bork, H. A. Proper, Surgery ai: Multimodal process mining and mixed reality for real-time surgical conformance checking and guidance, ZEUS 2025 (2025) 48.
[27] M. Fallon, J.-R. Rehse, A. Heinzl, Human behavior mining: A framework for theorizing about mhealth behavior using digital trace data, Journal of the Association for Information Systems 26 (2025) 633–667.
[28] R. Arablouei, Z. Wang, G. J. Bishop-Hurley, J. Liu, Multimodal sensor data fusion for in-situ classification of animal behavior using accelerometry and gnss data, Smart Agricultural Technology 4 (2023) 100163.
[29] S. Rinderle-Ma, J. Mangler, Process automation and process mining in manufacturing, in: International conference on business process management, Springer, 2021, pp. 3–14.
[30] A. Roitberg, N. Somani, A. Perzylo, M. Rickert, A. Knoll, Multimodal human activity recognition for industrial manufacturing processes in robotic workcells, in: Proceedings of the 2015 ACM on international conference on multimodal interaction, 2015, pp. 259–266.
[31] S. Knoch, S. Ponpathirkoottam, T. Schwartz, Video-to-model: Unsupervised trace extraction from videos for process discovery and conformance checking in manual assembly, in: International Conference on Business Process Management, Springer, 2020, pp. 291–308.
[32] W. Chanpuypetch, T. Supeekit, J. Niemsakul, Iot-based business process management for temperature-controlled logistics system of laboratory specimens., in: ECMS, 2023, pp. 359–365.
[33] Z. Kozhamkulova, L. Kuntunova, S. Amanzholova, A. Bizhanova, M. Vorogushina, A. Kuparova, M. Maikotov, E. Nurlybayeva, Development of intellectual decision making system for logistic business process management., International Journal of Advanced Computer Science & Applications 15 (2024).
[34] P. Valderas, V. Torres, E. Serral, Towards an interdisciplinary development of iot-enhanced business processes: P. valderas et al., Business & Information Systems Engineering 65 (2023) 25–48.
[35] L. Malburg, J. Grüger, R. Bergmann, An iot-enriched event log for process mining in smart factories, arXiv preprint arXiv:2209.02702 (2022).