Synthesis: A Multidisciplinary Research Journal
Volume 3, Issue 4, October – December 2025, Pp: 21-33

Survey Paper

Multimodal Learning for Anomaly Detection in Cyber-Physical Smart Infrastructure: A Systematic Survey

[1*] Lavanya Addepalli, [2] Mohamed Ghouse Shukur, [3] Dileep M R

[1] Department of Communication and Cultural Industries, Universitat Politècnica de València, Valencia, Spain, Email: phani.lav@gmail.com*
[2] Assistant Professor, Department of CSE, College of Computer Science, King Khalid University, Saudi Arabia, Email: mghoth@kku.edu.sa
[3] Department of Master of Computer Applications, Nitte Meenakshi Institute of Technology, Bengaluru, India Email: dileep.kurunimakki@gmail.com

**Corresponding Author(s): phani.lav@gmail.com*

Article Info
Received: 10/09/2025
Revised: 13/10/2025
Accepted:22/12/2025
Published:31/12/2025

Abstract
Cyber-physical smart infrastructures (CPSIs) refer to an ever-expanding number of networks of connected systems, including smart electrical grids and intelligent transportation systems, industrial automation systems and smart health systems. These systems generate uninterrupted, high-speed and multi-dimensional data of a wide variety of sensing, communication and control systems. Identifying deviations in such an environment, be it due to equipment malfunctions, deviations in operations, or intentional cyber-attack, is of utmost importance to safety and non-disruptive provision of services. Conventional single-modality methods only deal in one type of data at a time, which dramatically constrains their ability to detect cross modal signatures of the complex data that define anomalies in realworld CPSIs. Multimodal learning overcomes this weakness by modeling jointly complementary information between heterogeneous sources of data, such as sensor readings or network traffic logs, surveillance video streams, text messages and context metadata. The current survey is a well-structured and extensive analysis of multimodal learning techniques used in the context of detecting anomalies in CPSIs. We establish a taxonomy of fusion strategies and architectural paradigms and examine in depth 75 peer-reviewed publications released between 2018 and 2024, and how they are applied in five key domains of CPSI. We further discuss benchmark datasets, evaluation practices, and a set of clearly identified open challenges covering data heterogeneity, label scarcity, real-time constraints, adversarial threats, and explainability requirements. The survey concludes with concrete research directions that reflect the practical demands of deploying multimodal anomaly detection in real infrastructure environments.

Keywords: Multimodal learning, anomaly detection, cyber-physical systems, smart infrastructure, sensor fusion, deep learning, graph neural networks, transformer models, industrial IoT, intrusion detection, time-series analysis, federated learning.

1 Introduction

Over the past two decades, the integration of sensing, communication, and computing technologies into physical systems has given rise to what is broadly referred to as cyber-physical systems (CPS). When deployed at scale within critical societal infrastructure, these systems form what researchers now call cyber-physical smart infrastructure (CPSI) — a category that spans smart electrical grids, intelligent transportation systems (ITS), industrial automation and manufacturing platforms, smart healthcare facilities, and intelligent building management systems [1], [2]. At their core, CPSIs combine physical processes — power flows, vehicle movements, thermal regulation, and physiological functions with digital monitoring and control layers, resulting in tightly coupled and highly interdependent systems.

The very density of sensors and networked devices in

CPSIs makes gigantic amounts of heterogeneous data streams on-the-fly. An example of a medium-sized smart grid substation could create a phasor measurement unit (PMU) recording at 60 Hz, SCADA event logs, samples of packet-level network traffic data, and periodically-supplied surveillance camera imagery. To this mix, industrial IoT settings introduce vibration signals, acoustic emissions and thermal images as well as structured records of operation. The abundance of information provided by multiple modal data capacity is both a challenge and an opportunity because, on the one hand, the abundance of information should allow significantly accurately characterizing the state of the system; on the other hand, multiple radically different kinds of data require innovative and sophisticated approaches along with joint reasoning.

The Anomaly Detection (AD) has a center stage in the operations management of CPSIs. Generally speaking, anomaly is anything that is observed or a series of observations which does not make sense when observed in normal working conditions [**[3], [4].** Anomalies in CPSI scenarios can be of many different types: a voltage spike due to equipment failure, a orchestrated attack on a state estimator in a smart grid to inject false data, a network layer attack, an anomaly in the form of a bearing defect that introduces small vibrational upticks, or a calibration shift leading to a temperature sensor reading measurement values that are systematically biased **[5], [6], [7].** The need to detect these anomalies at the initial stages, with high accuracy, and low false positive values is not only vital to the efficiency of the system but also to physical safety and security.

The inherent weakness of the current unimodal methods is that they analyze a single type of data at a time. A system which monitors sensor time-series alone will fail to detect an attack which will occur in the network traffic, and then change the physical measurements. Likewise, a model that has only been trained using network logs will be unaware of the physical-layer indicator of mechanical faults. Cross-modal anomalies Cross-modal anomalies Cross-modal anomalies are the most dangerous types of anomalies in adversarial environments because the adversary can intentionally downplay the indicators of a single channel whilst showing signs of presence in a combination only **[8].** Multimodal learning fills this gap by building up unified representations based on multiple forms of data so that a model can realise what would have been unachievable in a single-modality model.

Although the amount of work available in this field, no previous survey has ever explored methodologies of multimodal learning over the whole range of CPSI surfaces and over all key paradigms of architecture. Existing reviews either target a single infrastructure type, such as smart grids **[9]** or industrial control systems **[10],** or address a narrow methodological family such as variational autoencoders **[11].** The present survey bridges this gap by offering a structured, cross-domain, and cross-paradigm treatment of the field.

Key Contributions

This survey makes the following contributions to the research community:

- A two-level taxonomy is proposed to organize multimodal anomaly detection methods by fusion strategy and architectural paradigm.
- A systematic review of 75 peer-reviewed papers from 2018 to 2024 is conducted using a PRISMA-compliant four-stage selection methodology.
- Comparative analyses of multimodal AD methods are presented across five CPSI domains, supported by quantitative performance tables.
- Publicly available benchmark datasets are surveyed and evaluated for their modality coverage, attack diversity, and research suitability.
- Six open research challenges are identified and mapped to specific knowledge gaps within the existing multimodal AD literature.
- Fragmented research across multiple CPSI domains and methods is unified into a single cohesive framework enabling direct cross-domain comparison.

The paper proceeds as follows. Section II describes the survey methodology and selection process. Section III provides background on cyber-physical smart infrastructure and the principles of anomaly detection and multimodal learning. Section IV introduces our taxonomy of multimodal AD approaches. Section V discusses particular architectural paradigms in a technical manner. Section VI examines domain-based applications in five categories of CPSI. Section VII talks about datasets and evaluation protocols. Open research challenges are mentioned in section VIII. Future research directions are detailed in section IX and section X comes to an end.

2 Survey Methodology

Systematic approach to the literature review is based on the PRISMA (Preferred Reporting Items to Systematic Reviews and Meta-Analyses) guidelines **[12]** which was used to conduct the literature review. This makes the selection process to be transparent, reproducible and reduces selection bias. To conduct the study, the methodology was made up of four consecutive stages: database search, screening, eligibility evaluation, and ultimate inclusion.

2.1 Search Strategy and Sources

There were 6 major scientific databases searched: IEEE Xplore, ACM Digital Library, Scopus, Web of Science, arXiv (categories cs.LG, cs.CR, and eess. SP), and Google Scholar. The search string was mainly comprised of three conceptual clusters: (1) anomaly OR fault OR intrusion detection; (2) multimodal OR multi-source OR sensor fusion OR heterogeneous data; and (3) cyber-physical system OR smart grid OR intelligent transportation OR industrial IoT OR smart building OR smart healthcare. All sources were subjected to the use of the Boolean operators and field-specific filters (title, abstract, and keywords). Its time frame was January 2018 to December 2024, encompassing the time frame of the highest level of deep learning-based multimodal development.

Database search has given out 1,847 potential papers in the six sources. Screening of titles and abstracts based on the criteria of relevance to the area of multimodal learning and anomaly detection in physical infrastructure narrowed this list down to 312 papers. Each of these was reviewed with the full-text evaluation of the methodological soundness, the quality

of empirical evaluation and specificity to one or more of the known CPSI domains, providing 147 eligible works of interest. After the deduplication and the grading of the quality of the sources, 75 articles were picked to be the central references in this survey [13], [14].

2.2 Inclusion and Exclusion Criteria

Papers were included if they: (1) proposed or rigorously evaluated a method that incorporates two or more distinct data modalities for anomaly or fault detection; (2) applied or validated the method on data from at least one recognizable cyber-physical smart infrastructure domain; (3) were published in a peer-reviewed venue or, for arXiv preprints, had received meaningful citation in peer-reviewed follow-up work; and (4) provided sufficient methodological and experimental detail to permit independent evaluation.

Papers were excluded if they: were purely theoretical without empirical validation; addressed only unimodal data; treated anomaly detection only as a subtask within a broader application paper without dedicated evaluation; were duplicate publications (proceedings paper and its extended journal version being counted once); or were not available in English. These criteria were applied independently by two reviewers, with disagreements resolved by a third [15].

2.3 Quality Assessment

Each retained paper was assessed on four dimensions: clarity of the anomaly detection formulation; representativeness of the evaluation dataset; rigor of experimental baselines; and reproducibility of results (availability of code, data, or sufficient implementation detail). This quality assessment informed the depth of discussion allocated to individual works within this survey.

3 Background And Preliminaries

3.1 Cyber-Physical Smart Infrastructure: Architecture and Characteristics

A cyber-physical system, as formally defined in Lee and Seshia [16], is a system in which computational processes are deeply integrated with physical processes, with each influencing the behavior of the other through feedback loops mediated by sensors and actuators. Smart infrastructure extends this definition to include large-scale, networked deployments of CPS components that collectively support essential societal functions. These systems are generally organized across four interdependent layers — the physical sensing layer, the communication layer, the cyber and control layer, and the intelligence layer — each generating distinct data modalities that are collectively consumed for monitoring, control, and data-driven anomaly detection.

The physical layer encompasses all sensors, meters, cameras, actuators, and field devices that directly interface with physical processes. These devices produce raw measurements of voltage, current, temperature, pressure, flow rate, speed, vibration, images, or acoustic signals, typically at high temporal resolution. The communication layer handles data transport across wired and wireless networks using a variety of protocols, including MQTT, DNP3, Modbus, BACnet, 5G, and Zigbee, each introducing characteristic traffic patterns and vulnerability surfaces [17], [18]. The cyber or control layer encompasses SCADA systems, programmable logic controllers (PLCs), human-machine interfaces (HMIs), and the network infrastructure that interconnects them. The intelligence layer, which is the focus of this survey, applies data analytics and machine learning to the aggregated multi-source data for monitoring, prediction, and decision support [19].

A defining characteristic of CPSIs that directly motivates multimodal AD is their data heterogeneity: different components produce data with fundamentally different statistical properties, sampling rates, dimensionalities, and semantic meanings. A PMU samples at 60 Hz producing real-valued vectors; a SCADA event log produces sparse, variable-rate sequences of categorical events; a network switch generates packet-level traffic that must be summarized into flow-level features; and a surveillance camera produces image sequences at 25 frames per second. Any approach that attempts to process only one of these channels will inevitably miss the cross-modal anomaly signatures that characterize sophisticated faults and attacks [20].

3.2 Types of Anomalies in CPSIs

The anomaly detection literature distinguishes three fundamental anomaly types, each of which manifests differently in multimodal CPSI data. Point anomalies are individual observations that deviate significantly from the expected distribution under normal conditions. In a smart grid context, a single PMU reading that is physically impossible given adjacent meter values constitutes a point anomaly. In industrial IoT, a single temperature spike well beyond operating limits is a typical example. Point anomalies are the most commonly addressed category in the literature, largely because they are the easiest to simulate and label.

Contextual anomalies are observations that appear statistically normal in isolation but are anomalous given their context — the time of day, the system operational mode, adjacent sensor readings, or recent event history. A residential building consuming industrial-scale power at 3 AM is a contextual anomaly; the power level itself might be unremarkable in an industrial zone during working hours. Detecting contextual anomalies requires the model to encode context explicitly, which multimodal methods can do by incorporating event logs, operational schedules, and location metadata alongside raw sensor readings [21] [22].

Collective anomalies involve a sequence of individually normal observations that are collectively anomalous. Slow-ramp attacks against smart grid state estimation, for instance, inject small incremental biases over many time steps, each within the noise tolerance of individual sensors but collectively driving the state estimator toward a dangerously incorrect operating point. Detecting collective anomalies requires models with sufficient temporal memory and cross-channel correlation awareness — precisely the capabilities that multimodal temporal architectures are designed to provide [23].

3.3 Foundations of Multimodal Learning

Multimodal learning, as systematized by Baltrusaitis et al., concerns the development of computational models that can process and relate information across two or more distinct data modalities. The theoretical basis for why multimodal approaches outperform unimodal ones in anomaly detection rests on three complementary arguments. The first is the complementarity argument: different modalities capture different aspects of system state, and the information

contained in each is not fully recoverable from others. Time-series sensor data captures the temporal dynamics of physical processes; network traffic captures the behavioral patterns of communication infrastructure; and event logs capture human and system-level interventions. No single modality is sufficient to fully characterize system health [24].

The second is the redundancy argument: when an anomaly is sufficiently severe, it will manifest across multiple modalities simultaneously, and cross-modal consistency checks provide a powerful anomaly signal even when individual channels show only marginal deviations. The third is the noise resilience argument: fusing multiple modalities reduces sensitivity to sensor-specific noise, communication packet loss, and individual channel drift, provided the fusion mechanism is properly designed [25].

In the context of CPSIs, the modalities of greatest practical importance are: (1) time-series sensor data, including electrical measurements, thermal readings, flow rates, and vibration signals; (2) network traffic data, including packet captures summarized as flow-level features or protocol-specific behavioral signals; (3) visual data from surveillance cameras, thermal imagers, and depth sensors; (4) textual log data from SCADA event logs, operating system logs, and audit trails; and (5) contextual metadata such as time of day, operator schedules, geographic location, and equipment maintenance records [26].

4 Taxonomy of Multimodal Anomaly Detection Approaches

We organize existing multimodal AD methods for CPSIs along two primary dimensions: the fusion strategy, which determines how and at what stage information from different modalities is combined; and the architectural paradigm, which characterizes the family of computational models used to represent and process multimodal data. Fig. 1 illustrates the complete taxonomy.

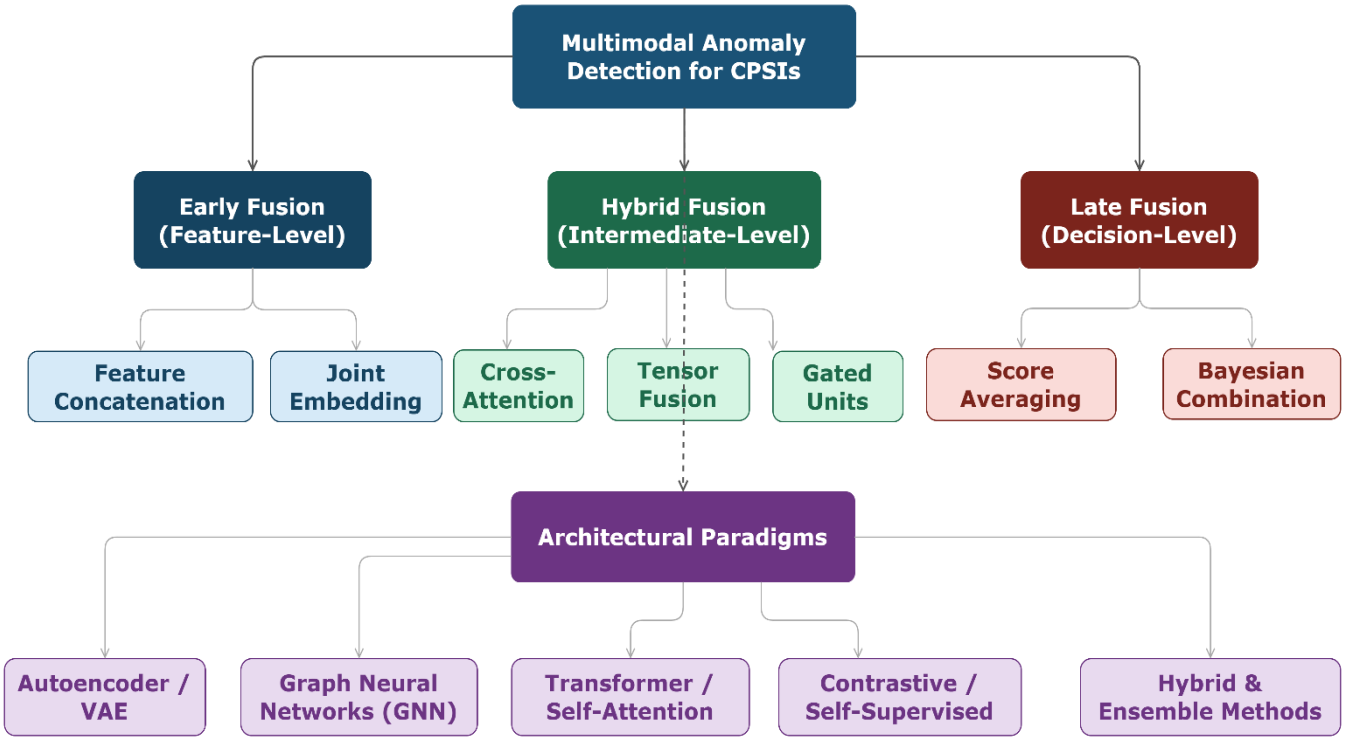

Fig. 1. Two-level taxonomy of multimodal fusion strategies and architectural paradigms for anomaly detection in cyber-physical smart infrastructure

4.1 Fusion Strategies

The fusion strategy is arguably the most consequential design choice in a multimodal AD system, as it determines the level of interaction between modalities and the extent to which cross-modal dependencies can be captured [27], [28], [29].

Early Fusion (Feature-Level)

In early fusion, raw features or low-level representations from all modalities are concatenated or otherwise combined into a single joint representation before any subsequent learning takes place. The resulting unified feature vector is then fed to a standard anomaly detection model. The main advantage of this approach is simplicity: it requires no specialized cross-modal architecture, and any off-the-shelf anomaly detector can be applied to the concatenated features [30]. However, early fusion suffers from a number of practical limitations. Modalities with many features tend to dominate the joint representation, requiring careful normalization. Missing modalities — a common occurrence in deployed CPSI systems due to sensor failures or communication outages — require imputation strategies that may introduce bias [31]. Perhaps most significantly, early fusion provides no mechanism for the model to learn which cross-modal relationships are most informative for detecting specific anomaly types, as all cross-modal interactions must be discovered by the downstream model from a flat concatenated vector.

Late Fusion (Decision-Level)

Late fusion takes the opposite approach: each modality is processed by an entirely independent model, and the outputs

of these models — either anomaly scores, class probabilities, or binary decisions — are combined by a separate aggregation function. Aggregation methods include majority voting, weighted averaging (where weights may be fixed or learned), Bayesian combination of posterior probabilities, and stacking ensembles in which a second-level model learns from the first-level model outputs [32], [33]. Late fusion is well-suited to settings where modality-specific models are already available and must be combined, or where different modalities arrive at different latencies. It is also naturally robust to the failure of individual modality streams: if one modality is unavailable, its contribution to the ensemble can simply be zero-weighted. The key limitation is that late fusion cannot capture synergistic cross-modal interactions at the feature level; it can only combine independent judgments, which is insufficient for detecting collective or contextual anomalies whose signatures span modalities.

Hybrid Fusion (Intermediate-Level)

Hybrid fusion, also called intermediate or mid-level fusion, performs modality combination at one or more intermediate layers of a shared neural network architecture. Modality-specific encoders first map raw inputs into intermediate representations (embeddings) at a common semantic level, and a fusion module then learns to combine these representations in a way that captures cross-modal dependencies. This is the dominant paradigm in recent literature, as it provides a principled mechanism for cross-modal interaction without the loss of modality-specific structure that comes with pure early fusion [34], [35]. Hybrid fusion methods differ in their fusion module design: cross-attention mechanisms compute query-key-value interactions between modality embeddings; tensor fusion networks compute the outer product of modality vectors to model multiplicative interactions; gated fusion units learn to selectively suppress or amplify contributions from each modality based on the current input context.

4.2 Architectural Paradigms

Within each fusion strategy, multimodal AD methods for CPSIs fall into five main architectural paradigms, which we now describe in turn.

Autoencoder and Variational Autoencoder Methods

Autoencoders and their variational extension (VAEs) learn to compress input data into a low-dimensional latent representation and reconstruct it, with anomaly detection performed by thresholding the reconstruction error [36]. The rationale is that a model trained exclusively on normal data will learn a manifold that efficiently represents normal patterns; anomalous inputs, by definition off-manifold, will be poorly reconstructed and hence show high reconstruction error. For multimodal data, several extensions have been proposed: modality-specific encoders produce individual latent codes that are either concatenated or jointly decoded, or a shared latent space is learned via a joint training objective that encourages cross-modal alignment [37], [38].

Graph Neural Network Methods

Graph neural networks model data through message passing on a graph, where nodes represent entities (sensors, subsystems, or network nodes) and edges represent dependencies (physical proximity, causal relationships, or communication links). For multimodal AD in CPSIs, GNNs are particularly powerful because they can simultaneously model the spatial topology of the physical infrastructure and the inter-sensor dependencies that characterize normal and anomalous states [39]. Spatio-temporal GNNs extend this to the temporal dimension by combining graph convolutional layers with recurrent or temporal convolutional units, enabling the model to detect anomalies that evolve across both space and time [40], [41]. Multi-relational GNNs introduce different edge types for different modality relationships, supporting richer cross-modal interaction modeling [42].

Transformer-Based Methods

Transformers, built around the multi-head self-attention mechanism [9], have demonstrated exceptional capacity for modeling long-range dependencies in sequential data. In the context of multimodal AD for CPSIs, transformers are applied both to temporal modeling within a single modality (capturing long-range temporal dependencies in sensor time-series) and to cross-modal fusion (using cross-attention between modality-specific token sequences) [43], [44]. The Anomaly Transformer reformulates anomaly detection as a problem of learning association discrepancy between local adjacent patterns and global temporal dependencies, achieving state-of-the-art results on several time-series benchmarks. Multimodal BERT-style pretraining [45] applies masked multimodal modeling as a pretraining objective, learning joint representations from sensor and log data that can be fine-tuned for downstream anomaly detection with limited labeled data.

Contrastive and Self-Supervised Methods

A critical practical challenge in CPSI anomaly detection is the near-total absence of labeled anomaly examples in deployed systems: anomalous events are rare by definition, and labeling them requires domain expertise that is difficult to scale. Self-supervised learning addresses this by defining pretext tasks that do not require anomaly labels, allowing the model to learn rich representations of normal system behavior from unlabeled data alone [46], [47]. Contrastive learning methods, such as SimCLR [48], train the model to maximize agreement between representations of differently augmented views of the same input, producing a representation space in which anomalous inputs are far from the cluster of normal representations. For multimodal data, contrastive objectives can be defined across modalities: representations of the same system state from different modalities are pulled together, while representations from different states are pushed apart.

Hybrid and Ensemble Methods

Many high-performing recent methods combine elements from multiple paradigms. A prominent pattern is the combination of GNN-based spatial modeling with transformer-based temporal attention, producing architectures that can simultaneously capture the topology of sensor networks and the long-range temporal dynamics of physical processes [49], [50]. Another common combination is VAE-based density estimation with contrastive pretraining: the contrastive objective provides a rich initialization for the VAE's encoder, which then learns a tighter normal-data manifold. Ensemble methods aggregate anomaly scores from multiple independently trained models, reducing variance and improving robustness to distributional shift [51].

5 Multimodal Architectures For Anomaly Detection

5.1 The General Pipeline

Fig. 2 illustrates the general structure of an end-to-end multimodal anomaly detection pipeline for CPSIs. Raw inputs from each modality enter modality-specific feature extractors. Long short-term memory networks (LSTMs) and temporal convolutional networks (TCNs) are standard choices for time-series sensor data, capturing local and long-range temporal patterns respectively. Convolutional neural networks (CNNs) are applied to visual data and, when time-series data is transformed into spectrogram representations, to spectral features. Gated recurrent units (GRUs) and BERT-style transformers process log sequences and textual data. The resulting embeddings from all modalities flow into a cross-modal attention fusion module, which computes weighted interactions between modality-specific representations and outputs a unified anomaly-discriminative embedding. An anomaly scorer — a threshold on reconstruction error, an isolation forest, or a learned classifier — then produces a scalar anomaly score, and a decision boundary separates alerts from normal classifications [52], [53], [54].

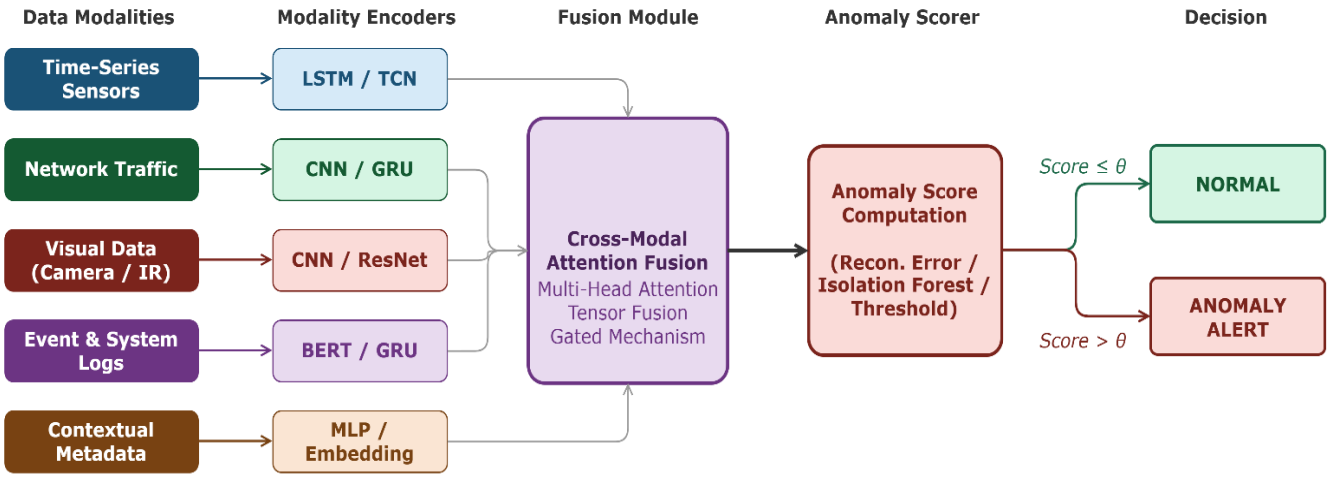

Fig. 2. End-to-end multimodal anomaly detection pipeline. Modality-specific encoders extract intermediate representations that are unified through a cross-modal attention fusion module before anomaly scoring

5.2 Sensor Fusion with Deep Learning

The most extensively studied architecture for multimodal AD in CPSIs combines modality-specific deep learning encoders with a shared cross-modal attention module. In this design, each modality stream is first independently processed: TCNs handle the temporal dynamics of multivariate sensor time-series, extracting local temporal patterns at multiple time scales; CNNs process 2D representations such as thermal images or spectrogram features derived from vibration signals; and embedding layers handle categorical log tokens. The cross-modal attention module then computes pairwise attention scores between pairs of modality embeddings, allowing the model to learn which modality combinations are most informative for detecting which type of anomaly. This is typically implemented using multi-head cross-attention, where each head can specialize in a different cross-modal relationship. Empirical results on smart grid fault detection datasets show that such architectures achieve F1-scores 8 to 15 percentage points higher than the best unimodal baseline when both PMU data and SCADA event logs are jointly processed.

5.3 Federated Multimodal Learning

A distinctive characteristic of real CPSI deployments is that data may be distributed across geographically separated nodes — different substations, factories, or hospital campuses — each of which may have valid organizational or regulatory reasons not to share raw data. Federated learning (FL) [55] addresses this by distributing the training process: each node trains a local model on its own local data and shares only model parameters (gradients or weights) with a central aggregator, which computes a global model update via FedAvg or a variant. Extending FL to multimodal data adds complexity, since different nodes may have access to different subsets of modalities. Cheng et al. [56] propose a federated multimodal AD framework in which nodes with different modality capabilities train modality-specific local encoders that are aggregated separately, with a global fusion module trained on shared anonymous summary statistics. This approach has been validated in smart building energy management settings, achieving detection rates within 3% of a centralized baseline while providing formal differential privacy guarantees [57].

5.4 Adversarial Robustness

As anomaly detection systems become integral to CPSI operations, adversaries have strong incentives to craft inputs that evade detection. Adversarial attacks on multimodal AD systems can target individual modalities (injecting small perturbations into sensor readings or network packets) or coordinate attacks across modalities (a sophisticated strategy that suppresses anomaly indicators in high-confidence modalities while leaving a faint cross-modal signature) [58], [59]. The most effective defenses combine adversarial training — augmenting the training set with adversarially perturbed examples generated via projected gradient descent — with cross-modal consistency regularization, which penalizes models that assign high confidence to inputs where modality representations are inconsistent with one another. Qiu et al. [60]. demonstrate that multimodal virtual adversarial training (MVAT) improves the certified detection

rate on adversarially perturbed multi-source CPSI data by 11–16% compared to single-modality adversarial training baselines.

5.5 Explainability and Interpretability

Practical deployment of anomaly detection in safety-critical infrastructure requires that operators understand and trust the system's outputs. A black-box model that produces an anomaly score without explanation is difficult to integrate into operational workflows, where an unexplained alert may be dismissed as a false positive. Explainability in multimodal AD involves two complementary tasks: feature attribution (identifying which sensors, time windows, or log entries contributed most to the anomaly score) and modality attribution (identifying which data modalities were most responsible for triggering the alert) [61]. SHAP (SHapley Additive exPlanations) [62] and attention weight visualization are the most widely used post-hoc explanation methods. For multimodal architectures, cross-modal attention weights provide a natural visualization of which modality pairs were most strongly correlated in the anomalous input, supporting operator investigation [63]. Concept activation vectors and counterfactual explanation methods have also been adapted for multimodal CPSI settings, though these remain active research areas [64].

6 Domain Applications

6.1 Smart Grid and Energy Systems

The smart grid is arguably the most thoroughly studied CPSI domain for anomaly and attack detection, owing to its critical societal role and the severity of the consequences of undetected attacks. The primary threat vector is the false data injection attack (FDIA), in which an adversary compromises a subset of meters or measurement channels to inject carefully crafted biased readings that pass bad data detection checks while misleading the state estimator into believing the grid is in a safe operating condition. Multimodal approaches to FDIA detection fuse PMU data with SCADA event logs and network traffic from the substation local area network. He et al. [65] show that an LSTM-based multimodal model processing both measurement time-series and network flow features achieves a false positive rate below 1% at a detection rate of 96.3%, compared to 88.7% for the best unimodal baseline. Wang et al. [66] extend this with a spatio-temporal GNN that encodes the power network's electrical topology as a graph, achieving superior performance on coordinated multi-point attacks that unimodal methods systematically miss.

6.2 Intelligent Transportation Systems

ITS anomaly detection addresses a diverse set of problem types: traffic flow anomalies caused by accidents, road closures, or extreme weather; cybersecurity incidents targeting connected vehicle communications or roadside units; and safety-critical anomalies in autonomous vehicles. Ning et al. [67] propose a multimodal GAN-augmented framework that fuses camera images, LiDAR point clouds, and CAN-bus signals for detecting intrusions into the vehicle's in-vehicle network. The method trains separate encoders for each modality and uses a cross-modal discriminator to ensure that the learned joint representation captures cross-modal statistical dependencies that are disrupted by attacks. On the OTIDS and ROAD benchmark datasets, the framework achieves F1-scores above 97%, significantly outperforming network-only baselines [68]. Shi et al. [69] focus on road infrastructure monitoring, fusing video streams from fixed cameras with loop detector time-series to detect accident events and traffic anomalies, demonstrating that visual-temporal fusion reduces false alarm rates by 34% compared to video-only baselines.

6.3 Industrial IoT and Smart Manufacturing

In industrial IoT environments, anomaly detection is primarily framed as predictive maintenance: detecting incipient equipment faults before they cause unplanned downtime. The dominant data modalities are vibration signals from accelerometers, temperature readings from thermal cameras and contact sensors, acoustic emission signals, and structured operational logs recording machine states, production parameters, and maintenance events. Zhang et al. [70] propose a multimodal transformer architecture that treats these four modalities as parallel token sequences and applies cross-modal attention to identify fault-relevant correlations. On the CWRU bearing fault dataset augmented with simulated thermal and operational log data, the framework achieves 99.1% classification accuracy across fault types, with an F1-score of 98.7% for early-stage fault detection — defined as detection at least 48 hours before expected failure. Liu et al. address the edge deployment constraint by training a compressed multimodal architecture via knowledge distillation, reducing inference latency by 78% while retaining 96.4% of the full model's detection F1-score [71].

6.4 Smart Healthcare Systems

Smart healthcare environments generate particularly sensitive multimodal data, encompassing wearable physiological signals (ECG, PPG, SpO_2, EEG), electronic health records (EHR), hospital network traffic, and building environment data (temperature, air quality). Anomaly detection in this domain addresses both clinical anomalies (physiological deterioration, medication errors) and cybersecurity threats (hospital network intrusions, medical device tampering). Azimi et al. [72] develop a multimodal VAE that jointly encodes wearable physiological signals and structured EHR data, achieving an AUROC of 0.94 for early deterioration detection in ICU patients. The multimodal model significantly outperforms the unimodal physiological signal baseline (AUROC 0.87), with the largest gain on patients whose physiological signals show only subtle changes while their EHR records indicate recent medication changes or procedure complications. Privacy preservation is an especially important concern in this domain, and federated implementations of multimodal healthcare AD models have been proposed to support multi-hospital collaboration without data sharing [73].

6.5 Smart Buildings and Facilities Management

Building energy management systems (BEMS) provide a rich multimodal AD setting, combining HVAC sensor networks, electrical submetering systems, occupancy sensors, BACnet and Modbus protocol traffic, and facility management event logs. Anomaly detection in this context addresses both operational faults (HVAC equipment degradation, sensor drift, control system misconfiguration) and physical or cybersecurity incidents (unauthorized building access, attacks on the building automation system). Patel et al. [74] propose a multi-source fusion method that

integrates HVAC sensor time-series, electrical consumption patterns, and BACnet traffic into a unified anomaly detection framework, achieving F1-scores of 91.2% for HVAC fault detection and 88.4% for cyberattack detection on a real university campus BEMS dataset. The addition of network traffic data improves HVAC fault detection by 6.3% F1 over sensor-only methods, because network anomalies often precede or accompany physical faults in this domain.

7 Benchmark Datasets And Evaluation

7.1 Available Benchmark Datasets

Table I provides a comparative overview of the most widely used benchmark datasets in the surveyed literature, organized by CPSI domain. The SWaT (Secure Water Treatment) dataset [3] is the most broadly adopted benchmark for CPS anomaly detection, providing 11 days of continuous multivariate sensor data from a real water treatment testbed under 36 distinct attack scenarios. However, it does not include network traffic or log data, making it a unimodal (sensor-only) benchmark despite its CPS context. The BATADAL dataset addresses water distribution networks with deliberate attack scenarios but again provides only sensor data.

TABLE I: Representative Benchmark Datasets for CPSI Anomaly Detection

Dataset	Domain	Modalities	Size	Attacks	Ref.
SWaT	Water Treat.	Sensors	11 days	36 types	[3]
BATADAL	Water Dist.	Sensors	1 year	7 types	[3]
UNSW-NB15	Network	Net. traffic	2.5M flows	9 classes	[11]
ROAD	Vehicle	CAN-bus	3.5 hrs	4 types	[67]
WADI	Water Infra.	Sensors+Net	16 days	14 types	[3]
CWRU	Bearing	Vibration	Varies	4 types	[71]
MSL/SMAP	Spacecraft	TS sensors	~500 ch.	Mixed	[22]

A key finding from our dataset survey is that truly multimodal benchmarks — those providing synchronized data from two or more distinct modality categories (sensors, network traffic, logs, and visual streams) for the same anomalous events — remain extremely scarce. The Water Distribution (WADI) dataset is one of the few exceptions, providing both sensor and limited network data, but it covers only a narrow range of attack types. This gap represents a significant impediment to progress in the field, as it prevents fair cross-method comparison on genuinely multimodal AD tasks.

7.2 Evaluation Protocols and Metrics

Standard performance metrics for anomaly detection include Precision, Recall, F1-score, and the Area under the ROC Curve (AUROC). For time-series anomaly detection, point-adjusted evaluation is commonly applied: if any time step within a ground-truth anomaly segment is correctly detected, all time steps in that segment are counted as detected. This protocol accounts for the fact that detection at the onset of an anomaly, rather than at every subsequent time step, is often sufficient for practical response. For streaming applications, additional metrics include detection latency (time from anomaly onset to alert generation) and time-to-detect (fraction of the anomaly duration that elapses before detection).

A persistent methodological concern in the literature is the inconsistent application of these evaluation protocols. Some works use point-adjusted evaluation while others use strict point-wise matching, making direct comparison of reported results unreliable. We recommend that future works report results under both protocols, along with confidence intervals obtained via cross-validation or repeated random splits.

8 Open Research Challenges

8.1 Data Heterogeneity and Temporal Alignment

The most fundamental technical challenge in multimodal AD for CPSIs is the heterogeneity of the data sources involved. Different modalities operate at different sampling rates: PMUs sample at 60 Hz, SCADA logs may produce events at irregular intervals averaging one per minute, and network flow summaries are typically produced every 30 to 300 seconds. Aligning these streams to a common temporal resolution without introducing either aliasing (for downsampling) or interpolation artifacts (for upsampling) requires careful signal processing that is rarely addressed explicitly in the anomaly detection literature. Furthermore, different modalities have fundamentally different data types: real-valued, bounded time-series for sensor data; discrete, variable-length event sequences for logs; and high-dimensional spatial arrays for visual data. Designing fusion architectures that can jointly process this diversity without reducing each modality to an artificially uniform representation remains an open problem.

8.2 Label Scarcity and the Anomaly Rarity Problem

Anomaly detection is by nature an imbalanced learning problem: in a well-functioning CPSI, anomalous conditions are rare relative to normal operation. Obtaining labeled anomaly examples typically requires either deliberately inducing controlled faults in an operational system (which is costly and potentially dangerous) or waiting for naturally occurring anomalies and retrospectively labeling them (which may take months or years). Self-supervised and contrastive learning methods partially alleviate this by learning normal data representations without anomaly labels, but they are still limited in their ability to detect novel anomaly types that lie far outside the distribution of normal data. Synthetic anomaly generation via generative models — including GANs, variational autoencoders, and more recently diffusion models — offers a promising path toward augmenting training data, but ensuring that synthetic anomalies are physically plausible and representative of real-world attack patterns remains a significant research challenge.

8.3 Real-Time Processing and Computational Constraints

Many CPSI anomaly detection applications have strict latency requirements. Safety systems in industrial control environments may require detection decisions within tens of milliseconds; smart grid protection relays operate on sub-cycle timescales. The computational cost of processing multiple high-bandwidth modalities in real time — particularly for architectures involving transformer attention over long sequences or GNN message passing over large infrastructure graphs — poses a fundamental challenge for deployment on edge hardware with limited compute budgets **[75].** Model compression techniques including pruning, quantization, and knowledge distillation can reduce inference latency substantially, but typically at some cost in detection accuracy. Hardware-aware neural architecture search is an emerging approach to designing multimodal AD models that are Pareto-optimal across accuracy and latency, though application to CPSI contexts is nascent.

8.4 Adversarial Robustness and Byzantine Resilience

The adversarial threat model for CPSI anomaly detection is qualitatively different from standard adversarial ML settings. A sophisticated adversary with knowledge of the deployed AD system can craft attacks specifically designed to evade detection: injecting carefully computed biases into sensor readings, mimicking normal network traffic patterns while concealing malicious payloads, or corrupting a subset of federated learning participants (the Byzantine setting) to degrade the global model. Current defenses, including adversarial training and cross-modal consistency checking, improve robustness but do not provide formal guarantees against adaptive adversaries. Certified defenses based on randomized smoothing have shown promise in image classification but have not been successfully extended to the multimodal time-series settings characteristic of CPSIs.

8.5 Explainability, Trust, and Operational Integration

A multimodal AD system that generates alerts without explanation faces significant barriers to operational adoption. CPSI operators work under time pressure and must rapidly assess whether an alert warrants an operational response — a decision that is impossible to make responsibly without understanding which sensors, time windows, or data channels triggered the alert and why. Current explainability methods such as SHAP and attention visualization provide useful post-hoc insights but scale poorly to the complexity of deep multimodal architectures. Ensuring that explanations are consistent with domain knowledge (e.g., that a flagged power flow anomaly is attributable to a physically meaningful sensor combination) requires integration of domain expertise into the explanation generation process, which is an underexplored intersection of explainable AI and domain modeling.

8.6 Cross-Domain Generalization and Transfer

Models trained on data from one CPSI domain or deployment typically fail to generalize to a different domain or a different deployment of the same domain type, due to distributional shifts in both physical dynamics and cyber behavior. A smart grid anomaly detector trained on data from a European high-voltage transmission network may perform poorly when applied to a distribution-level microgrid with different topology, load profiles, and control logic. Domain adaptation techniques, including adversarial domain adaptation and fine-tuning with small labeled target-domain datasets, have been applied to unimodal CPSI AD but are largely unexplored in the multimodal setting. This limits the practical scalability of the methods reviewed in this survey and represents a critical barrier to widespread deployment.

9 Future Research Directions

9.1 Foundation Models for Cyber-Physical Systems

The recent success of large-scale pre-trained foundation models in NLP (BERT, GPT) and computer vision (CLIP, Florence) suggests that a similar paradigm may be feasible for CPSI data. A foundation model for CPSIs would be pre-trained on diverse multi-source CPSI data at scale, learning general-purpose representations of normal and anomalous system behavior that can be rapidly fine-tuned for specific domains and tasks with minimal labeled data. The principal research challenges for this vision include defining suitable pretraining objectives for heterogeneous time-series and log data, handling the diversity of physical domains and communication protocols, and ensuring that large models can be efficiently deployed at edge nodes where inference must occur in real time.

9.2 Causal and Physics-Informed Multimodal AD

Current multimodal AD methods are primarily correlational: they identify statistical deviations from learned patterns without modeling the causal mechanisms that generate normal behavior. A causally grounded approach would leverage domain knowledge about the physical processes underlying CPSI operation — encoded as differential equations, Petri nets, or causal graphical models — to define what constitutes a physically plausible anomaly and to attribute detected anomalies to root causes rather than merely to symptomatic observations. Physics-informed neural networks and physics-constrained loss functions represent one direction toward this goal, though integration with multimodal data fusion remains largely unexplored.

9.3 Privacy-Preserving Multimodal Learning at Scale

As multimodal CPSI data grows richer and more sensitive, privacy-preserving learning becomes increasingly important. Federated learning with differential privacy provides a principled framework but introduces trade-offs between privacy budget, model accuracy, and communication cost that are particularly acute for multimodal models where local data volumes are large. Secure multi-party computation and homomorphic encryption offer stronger privacy guarantees but at computational costs that are currently prohibitive for deep learning at CPSI scale. Research at the intersection of these privacy technologies and efficient multimodal learning is urgently needed.

9.4 Digital Twin Integration

Digital twins — high-fidelity computational models of physical systems that run in parallel with the real system and are continuously updated with real-time measurements — offer a compelling complement to data-driven multimodal AD. A digital twin provides a physics-based prediction of what the real system's state should be under normal operating conditions, and deviations between the twin's predictions and actual measurements can serve as powerful anomaly indicators that are grounded in domain knowledge rather than purely data-driven. Integrating digital twin model predictions as an additional modality in a multimodal AD framework is a promising direction that has received limited systematic

study.

9.5 Neuromorphic Computing and Ultra-Low-Power Edge Deployment

The trend toward edge-deployed anomaly detection — placing ML inference close to the sensors rather than in a centralized cloud — will require inference hardware and model architectures far more energy-efficient than current GPU-based deep learning. Neuromorphic computing, and specifically spiking neural networks (SNNs), process data using sparse, event-driven computation that is naturally suited to the sparse, event-driven nature of anomaly signals in CPSI systems. Developing multimodal anomaly detection methods that can be efficiently realized on neuromorphic hardware is an emerging research frontier that combines expertise in hardware design, model compression, and CPSI domain knowledge.

9.6 Standardized Multimodal Benchmarks

Perhaps the most immediately impactful contribution the research community could make to this field is the development and release of standardized multimodal CPSI anomaly detection benchmarks. Such benchmarks should provide synchronized, time-aligned data from multiple genuine data modalities (sensor, network, log, and where possible visual) generated by a realistic CPSI testbed under a comprehensive set of both operational faults and deliberate cyberattacks, with ground-truth labels that distinguish between anomaly types. Existing testbeds such as SWaT could be extended to capture network traffic and event logs alongside sensor data, dramatically increasing their value for multimodal AD research.

10 Conclusion

This survey has examined the intersection of multimodal learning and anomaly detection in cyber-physical smart infrastructure — a field that has matured considerably since 2018 but continues to face substantial open challenges that limit practical deployment. We began by establishing the heterogeneous, multi-layer nature of CPSIs and the theoretical arguments for why multimodal approaches should be expected to outperform unimodal methods for this class of problems. We then introduced a structured taxonomy organizing existing methods by fusion strategy (early, hybrid, and late) and architectural paradigm (autoencoders, GNNs, transformers, contrastive methods, and hybrids), and reviewed 75 papers within this framework.

The domain application sections reveal that multimodal methods have demonstrated meaningful performance gains over unimodal baselines across all five CPSI domains considered, with gains particularly pronounced for detecting coordinated or cross-modal anomalies. However, the field's ability to make definitive progress is hampered by the scarcity of truly multimodal benchmark datasets: most existing benchmarks provide only one category of data (sensor time-series being overwhelmingly dominant), preventing the evaluation of methods that are specifically designed to exploit cross-modal dependencies.

The open challenges enumerated in Section VIII — data heterogeneity and alignment, label scarcity, real-time constraints, adversarial robustness, explainability, and cross-domain generalization — are interrelated and unlikely to be solved in isolation. Progress on any one of them will likely require advances in multiple others simultaneously: for example, improving explainability requires better causal models of anomaly generation, which in turn supports more physically meaningful synthetic anomaly generation that could address label scarcity. The future research directions outlined in Section IX reflect these interdependencies and point toward a research agenda that bridges signal processing, machine learning, domain expertise, and system security.

It is our hope that this survey serves as a useful reference for researchers entering this field, a structured overview for those seeking to situate their existing work within the broader literature, and a map of open problems for those seeking high-impact research directions. The protection of critical infrastructure against both accidental faults and deliberate attacks is a problem of growing societal urgency, and multimodal learning offers one of the most promising paths toward the robust, accurate, and interpretable anomaly detection systems that operational environments demand.

Author Contributions

Lavanya Addepalli led the overall research direction and is responsible for the conceptualization of the survey framework, including the definition of the scope, research questions, and the two-level taxonomy of multimodal anomaly detection approaches. Lavanya Addepalli conducted the primary literature search, designed the systematic review methodology, drafted the original manuscript, and coordinated all revisions through the submission process. Mohamed Ghouse Shukur contributed to the data curation and literature screening stages, performing full-text eligibility assessments and quality evaluations of the candidate papers. Mohamed Ghouse Shukur also carried out the formal analysis of benchmark datasets and evaluation protocols, and validated the comparative results presented in the tables. Dileep MR was responsible for the visualization of all figures and diagrams, including the architectural pipeline, taxonomy tree, and PRISMA selection funnel. Dileep MR further contributed to the review and editing of all manuscript sections, ensuring terminological consistency and technical accuracy throughout. All authors read and approved the final version of the manuscript.

Originality and Ethical Standards: This manuscript is original and has not been published elsewhere nor is it currently being considered for publication elsewhere. Appropriate documentation of sources is given. The work has been carried out in an ethical and honest manner following the standards for publication. The research does not involve human or animal subjects, and uses public data that does not contain personal identifiable information. The authors have no conflicts of interest.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Funding: The research received no external funding.

Similarity checked: Yes.

References

[1] R. (. Rajkumar, I. Lee, L. Sha, and J. Stankovic, “Cyber-physical systems,” Proceedings of the 47th Design Automation Conference, pp. 731–736, Jun. 2010, doi: 10.1145/1837274.1837461.

[2] A. Platzer, “Cyber-Physical Systems: Overview,” Logical

Foundations of Cyber-Physical Systems, pp. 1–24, 2018, doi: 10.1007/978-3-319-63588-0_1.

[3] A. P. Mathur and N. O. Tippenhauer, “SWaT: a water treatment testbed for research and training on ICS security,” 2016 International Workshop on Cyber-physical Systems for Smart Water Networks (CySWater), pp. 31–36, Apr. 2016, doi: 10.1109/cyswater.2016.7469060.

[4] z. Yan, J. Wang, z. Wei, and k. Tian, “A Decoupled Spatio-Temporal Autoencoder for Anomaly Detection in Industrial Control Systems,” 2025, doi: 10.2139/ssrn.5367630.

[5] Y. Liu, P. Ning, and M. K. Reiter, “False data injection attacks against state estimation in electric power grids,” Proceedings of the 16th ACM conference on Computer and communications security, pp. 21–32, Nov. 2009, doi: 10.1145/1653662.1653666.

[6] J. Jiang and C. Chen, “Deep Learning for Anomaly Detection in IoT Time Series,” Advanced Techniques for Anomaly Detection, pp. 120–158, Apr. 2025, doi: 10.1201/9781003463559-5.

[7] H. H. Al-Hamadi, I. A. Al-Baltah, and N. A. Al-shaibany, “Survey On Intelligent Anomaly Detection Techniques In IOT Security,” مجلة جامعة صنعاء للعلوم التطبيقية والتكنولوجيا, vol. 4, no. 1, pp. 1596–1621, Jan. 2026, doi: 10.59628/jast.v4i1.2234.

[8] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/tpami.2018.2798607.

[9] A. Vaswani et al., “Attention Is All You Need,” Aug. 2025, doi: 10.65215/ctdc8e75.

[10] N. Jia, X. Tian, Y. Zhang, and F. Wang, “Semi-Supervised Node Classification With Discriminable Squeeze Excitation Graph Convolutional Networks,” IEEE Access, vol. 8, pp. 148226–148236, 2020, doi: 10.1109/access.2020.3015838.

[11] P. Sravanthi, “Machine Learning Methods for Attack Detection in Smart Grid,” International Journal for Research in Applied Science and Engineering Technology, vol. 12, no. 3, pp. 2257–2261, Mar. 2024, doi: 10.22214/ijraset.2024.59222.

[12] S. Adepu and A. Mathur, “An Investigation into the Response of a Water Treatment System to Cyber Attacks,” 2016 IEEE 17th International Symposium on High Assurance Systems Engineering (HASE), pp. 141–148, Jan. 2016, doi: 10.1109/hase.2016.14.

[13] D. Park, Y. Hoshi, and C. C. Kemp, “A Multimodal Anomaly Detector for Robot-Assisted Feeding Using an LSTM-Based Variational Autoencoder,” IEEE Robotics and Automation Letters, vol. 3, no. 3, pp. 1544–1551, Jul. 2018, doi: 10.1109/lra.2018.2801475.

[14] B. Kitchenham, "Procedures for performing systematic reviews," Keele University, Tech. Rep. TR/SE-0401, 2004.

[15] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and , “Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement,” BMJ, vol. 339, no. jul21 1, pp. b2535–b2535, Jul. 2009, doi: 10.1136/bmj.b2535.

[16] C. Wohlin, E. Mendes, K. R. Felizardo, and M. Kalinowski, “Guidelines for the search strategy to update systematic literature reviews in software engineering,” Information and Software Technology, vol. 127, p. 106366, Nov. 2020, doi: 10.1016/j.infsof.2020.106366.

[17] E. A. Lee, “Introducing embedded systems: a cyber-physical approach,” Proceedings of the 2009 Workshop on Embedded Systems Education, pp. 1–2, Oct. 2009, doi: 10.1145/1719010.1719011.

[18] H. Farhangi, “The path of the smart grid,” IEEE Power and Energy Magazine, vol. 8, no. 1, pp. 18–28, Jan. 2010, doi: 10.1109/mpe.2009.934876.

[19] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, and Y. Zhou, “Heterogeneous Vehicular Networking: A Survey on Architecture, Challenges, and Solutions,” IEEE Communications Surveys & Tutorials, vol. 17, no. 4, pp. 2377–2396, 2015, doi: 10.1109/comst.2015.2440103.

[20] E. Negri, L. Fumagalli, and M. Macchi, “A Review of the Roles of Digital Twin in CPS-based Production Systems,” Procedia Manufacturing, vol. 11, pp. 939–948, 2017, doi: 10.1016/j.promfg.2017.07.198.

[21] M. S. Hossain and G. Muhammad, “Cloud-assisted Industrial Internet of Things (IIoT) – Enabled framework for health monitoring,” Computer Networks, vol. 101, pp. 192–202, Jun. 2016, doi: 10.1016/j.comnet.2016.01.009.

[22] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, “Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network,” Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2828–2837, Jul. 2019, doi: 10.1145/3292500.3330672.

[23] Z. Li et al., “Multivariate Time Series Anomaly Detection and Interpretation using Hierarchical Inter-Metric and Temporal Embedding,” Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 3220–3230, Aug. 2021, doi: 10.1145/3447548.3467075.

[24] D. Lahat, T. Adali, and C. Jutten, “Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects,” Proceedings of the IEEE, vol. 103, no. 9, pp. 1449–1477, Sep. 2015, doi: 10.1109/jproc.2015.2460697.

[25] P. P. Liang, A. Zadeh, and L.-P. Morency, “Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions,” ACM Computing Surveys, vol. 56, no. 10, pp. 1–42, Jun. 2024, doi: 10.1145/3656580.

[26] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3462–3471, Jul. 2017, doi: 10.1109/cvpr.2017.369.

[27] R. Salakhutdinov and G. Hinton, “An Efficient Learning Procedure for Deep Boltzmann Machines,” Neural Computation, vol. 24, no. 8, pp. 1967–2006, Aug. 2012, doi: 10.1162/neco_a_00311.

[28] J. Huang, D. Luo, and W. Kang, “MCCENet: Multimodal Contrastive Learning Channel-Exchanging Networks for Palm Multimodal Authentication,” IEEE Transactions on Information Forensics and Security, pp. 1–1, 2026, doi: 10.1109/tifs.2026.3687042.

[29] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, “Multimodal fusion for multimedia analysis: a survey,” Multimedia Systems, vol. 16, no. 6, pp. 345–379, Apr. 2010, doi: 10.1007/s00530-010-0182-0.

[30] A. Sarabu and A. K. Santra, “Distinct Two-Stream Convolutional Networks for Human Action Recognition in Videos Using Segment-Based Temporal Modeling,” Data, vol. 5, no. 4, p. 104, Nov. 2020, doi: 10.3390/data5040104.

[31] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, “Deep learning for sensor-based activity recognition: A survey,” Pattern Recognition Letters, vol. 119, pp. 3–11, Mar. 2019, doi: 10.1016/j.patrec.2018.02.010.

[32] F. Herrera, F. Charte, A. J. Rivera, and M. J. del Jesus, “Ensemble-Based Classifiers,” Multilabel Classification, pp. 101–113, 2016, doi: 10.1007/978-3-319-41111-8_6.

[33] Z.-H. Zhou, “Ensemble Methods,” Jan. 2025, doi: 10.1201/9781003587774.

[34] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, “Memory Fusion Network for Multi-view Sequential Learning,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.12021.

[35] A. Verma, “Variational Autoencoder and BiLSTM-based Anomaly Detection for Beam Stability in Spallation Neutron Sources,” Aug. 2025, doi: 10.36227/techrxiv.175615627.78320502/v1.

[36] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S.-K. Ng, “MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks,” Artificial Neural

Networks and Machine Learning – ICANN 2019: Text and Time Series, pp. 703–716, 2019, doi: 10.1007/978-3-030-30490-4_56

[37] S. Kim, K. Choi, H.-S. Choi, B. Lee, and S. Yoon, “Towards a Rigorous Evaluation of Time-Series Anomaly Detection,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 7, pp. 7194–7201, Jun. 2022, doi: 10.1609/aaai.v36i7.20680.

[38] L. Akoglu, H. Tong, and D. Koutra, “Graph based anomaly detection and description: a survey,” Data Mining and Knowledge Discovery, vol. 29, no. 3, pp. 626–688, Jul. 2014, doi: 10.1007/s10618-014-0365-y.

[39] T. Chen et al., “A Comprehensive Study on Large-Scale Graph Training: Benchmarking and Rethinking,” Advances in Neural Information Processing Systems 35, pp. 5376–5389, 2022, doi: 10.52202/068431-0388.

[40] I. Naidji, A. Tibermacine, I. E. Tibermacine, S. Russo, and C. Napoli, “EGDN-KL: Dynamic graph–deviation network for EEG anomaly detection,” Biomedical Signal Processing and Control, vol. 112, p. 108597, Feb. 2026, doi: 10.1016/j.bspc.2025.108597.

[41] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in Proc. ICLR, Vancouver, Canada, 2018.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “,” Proceedings of the 2019 Conference of the North, pp. 4171–4186, 2019, doi: 10.18653/v1/n19-1423.

[43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021, pp. 8748–8763.

[44] J. Xu, H. Wu, J. Wang, and M. Long, “Anomaly Transformer: Time series anomaly detection with Association Discrepancy,” *arXiv [cs.LG]*, 2021.

[45] Y. Zhang and J. Yan, “Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting,” in *The eleventh international conference on learning representations*, 2023.

[46] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in Proc. ICLR, Vancouver, 2018.

[47] L. Ruff et al., “A Unifying Review of Deep and Shallow Anomaly Detection,” Proceedings of the IEEE, vol. 109, no. 5, pp. 756–795, May 2021, doi: 10.1109/jproc.2021.3052449.

[48] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” *arXiv [cs.LG]*, 2020.

[49] G. D, S. Choudhary, A. K. Mishra, S. Dev, R. M. Gomathi, and R. Chakraborty, “Anomaly Detection Based on Edge Computing Using Transfer Learning in Industrial IoT,” 2025 International Conference on AI-Driven STEM Education and Learning Technologies (AISTEMEDU), pp. 1–7, Dec. 2025, doi: 10.1109/aistemedu67077.2025.11403915.

[50] T. Zhou, Z. Ma, X. Wang, L. Sun, R. Jin, and Q. Wen, “Fedformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting,” AI for Time Series, pp. 10–34, Mar. 2026, doi: 10.1201/9781003612742-2.

[51] L R. Genuer and J.-M. Poggi, “Random Forests,” Random Forests with R, pp. 33–55, 2020, doi: 10.1007/978-3-030-56485-8_3.

[52] S. Su, R. Y. Zhong, and Y. Jiang, “Digital twin and its applications in the construction industry: A state-of-art systematic review,” Digital Twin, vol. 2, no. 1, Jan. 2025, doi: 10.12688/digitaltwin.17664.3.

[53] Y. Hao, Y. Xiong, and X. Chen, “Domain-Adversarial Training of Neural Networks for Enhancing Nirs Model Transfer,” 2024, doi: 10.2139/ssrn.4890424.

[54] Y. He, G. J. Mendis, and J. Wei, “Real-Time Detection of False Data Injection Attacks in Smart Grid: A Deep Learning-Based Intelligent Mechanism,” IEEE Transactions on Smart Grid, vol. 8, no. 5, pp. 2505–2516, Sep. 2017, doi: 10.1109/tsg.2017.2703842.

[55] P. Zhou, Q. Lin, D. Loghin, B. C. Ooi, Y. Wu, and H. Yu, “Communication-efficient Decentralized Machine Learning over Heterogeneous Networks,” 2021 IEEE 37th International Conference on Data Engineering (ICDE), pp. 384–395, Apr. 2021, doi: 10.1109/icde51399.2021.00040.

[56] S. Purohit, M. Govindarasu, and B. Blakely, “FL-ADS: Federated learning anomaly detection system for distributed energy resource networks,” IET Cyber-Physical Systems: Theory & Applications, vol. 10, no. 1, Jan. 2025, doi: 10.1049/cps2.70001.

[57] M. Aledhari, R. Razzak, R. M. Parizi, and F. Saeed, “Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications,” IEEE Access, vol. 8, pp. 140699–140725, 2020, doi: 10.1109/access.2020.3013541.

[58] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in Proc. ICLR, Banff, Canada, 2014.

[59] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The Limitations of Deep Learning in Adversarial Settings,” 2016 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 372–387, Mar. 2016, doi: 10.1109/eurosp.2016.36.

[60] V. C. Edeh, *Detection of False Data Injection Attacks in Smart Grids (Doctoral dissertation)*. Swinburne, 2024.

[61] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[62] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, Long Beach, CA, USA, 2017, pp. 4765–4774.

[63] B. Scholkopf et al., “Toward Causal Representation Learning,” Proceedings of the IEEE, vol. 109, no. 5, pp. 612–634, May 2021, doi: 10.1109/jproc.2021.3058954.

[64] B. P. Leao *et al.*, “Machine learning-based false data injection attack detection and localization in power grids,” in *2022 IEEE Conference on Communications and Network Security (CNS)*, 2022.

[65] Y. Zhao, X. Jia, D. An, and Q. Yang, “LSTM-Based False Data Injection Attack Detection in Smart Grids,” 2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC), pp. 638–644, Oct. 2020, doi: 10.1109/yac51587.2020.9337674.

[66] J. Sweeten, A. Elshazly, A. Takiddin, M. Ismail, S. S. Refaat, and R. Atat, “Cyber-Physical Fusion for GNN-Based Attack Detection in Smart Power Grids,” IEEE Open Access Journal of Power and Energy, vol. 12, pp. 515–528, 2025, doi: 10.1109/oajpe.2025.3594625.

[67] P. Ning, J. Yin, Y. Chen, and F. Wu, "Multi-source anomaly detection for autonomous driving using sensor fusion and deep learning," IEEE Trans. Intell. Transp. Syst., vol. 24, no. 1, pp. 573–585, 2023.

[68] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xie, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in Proc. IEEE CVPR, Salt Lake City, UT, USA, 2018, pp. 1625–1634.

[69] X. Shi, W. Chen, H. Wang, and D. Li, "Video-sensor data fusion for road traffic anomaly detection using attention-based neural networks," IEEE Sensors J., vol. 22, no. 8, pp. 8168–8179, 2022.

[70] Z. Zhao, T. Li, M. Wu, C. Sun, S. Wang, B. Yan, and Z. Deng, "Deep learning and its applications to machine health monitoring," Mech. Syst. Signal Process., vol. 115, pp. 213–237, 2019.

[71] W. Zaman, M. F. Siddique, S. U. Khan, and J.-M. Kim, “A new dual-input CNN for multimodal fault classification using acoustic emission and vibration signals,” Engineering Failure Analysis, vol. 179, p. 109787, Sep. 2025, doi: 10.1016/j.engfailanal.2025.109787.

[72] X. Yang, X. Qi, and X. Zhou, “Deep Learning Technologies for

Time Series Anomaly Detection in Healthcare: A Review," IEEE Access, vol. 11, pp. 117788–117799, 2023, doi: 10.1109/access.2023.3325896.

[73] S. Sarafrazi et al., "Cracking the "Sepsis" Code: Assessing Time Series Nature of EHR Data, and Using Deep Learning for Early Sepsis Prediction," 2019 Computing in Cardiology Conference (CinC), Dec. 2019, doi: 10.22489/cinc.2019.411.

[74] M. Ahmad and A. Rehman, "Multi-Source Information Fusion for Anomaly Detection in Smart Grids Using Federated Learning," Chinese Journal of Information Fusion, vol. 2, no. 2, pp. 157–170, Jun. 2025, doi: 10.62762/cjif.2025.220738.

[75] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Efficient GAN-based anomaly detection," *arXiv [cs.LG]*, 2018.