Comparative Evaluation of Multimodal Fusion Architectures for Cyber Physical System Anomaly Classification

Lukmon Rasaq[1], Om Prakash Yadav[2], Madhuri Siddula[3], Joao Paulo Jacomini Prioli [4]

[1,2,3,4]*North Carolina A&T State University, Greensboro, NC, 27411, USA*

{larasaq@aggies.ncat.edu, oyadav@ncat.edu, msiddula@ncat.edu, jjacominiprioli@ncat.edu}

ABSTRACT

Smart manufacturing systems generate heterogeneous data from programmable logic controllers (PLCs), industrial networks, and system logs, providing complementary information for cyber-physical system health monitoring. However, few studies have systematically compared multimodal fusion architectures for industrial anomaly classification. This paper evaluates four architectures, Fusion MLP, Fusion 1D CNN, Fusion CNN LSTM, and Cross Modal Attention Fusion, using synchronized PLC telemetry, network traffic, and system log data under an identical experimental protocol. Unimodal and classical machine learning baselines are also evaluated to assess the contribution of multimodal fusion. Results show that all fusion architectures achieve near-perfect classification performance. However, unimodal and temporal evaluations reveal that the benchmark is highly separable, with PLC telemetry alone achieving 99.93% accuracy and 99.68% macro F1 under temporal evaluation. Confusion matrices, t-SNE visualization, and attention allocation analysis further examine model behavior and modality contributions. The findings show that multimodal fusion effectively integrates heterogeneous industrial data, while increased architectural complexity does not necessarily improve classification performance when individual modalities contain highly discriminative signals. These results highlight the importance of unimodal baselines, temporal robustness evaluation, and interpretable multimodal analysis in assessing industrial anomaly detection systems.

1. INTRODUCTION

The rapid adoption of Industry 4.0 technologies has transformed modern manufacturing into highly interconnected cyber–physical environments that continuously generate heterogeneous data from programmable logic controllers (PLCs), industrial communication networks, machine sensors, and system logs. Integrating these complementary information sources enables more comprehensive monitoring of manufacturing assets and supports intelligent anomaly detection, predictive maintenance, and autonomous decision-making (Cheng Ren, Ming Li, Cailian Chen, Xinping Guan, and George Q. Huang 2026, Jeongheun Kang, Jiwon Lee, and Jongpil Jeong 2025). Consequently, multimodal anomaly detection has emerged as a promising research direction for improving the reliability, robustness, and operational resilience of smart manufacturing systems.

Recent advances have primarily focused on multimodal industrial anomaly detection using RGB images and 3D point clouds. (Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, Chengjie Wang, 2023) proposed Multi-3D-Memory (M3DM), which introduced a hybrid fusion strategy combining feature-level interaction and decision-level fusion to reduce interference between modalities and improve anomaly detection performance. Building upon this direction, (Hao Cheng, Jiaxiang Luo, and Xianyong Zhang, 2025) proposed a uni-modal and cross-modal fusion framework that explicitly models both intra-modal and inter-modal relationships, demonstrating superior anomaly localization over conventional feature concatenation approaches. More recently, (Jiaxun Wang, Yanchang Niu, Biqing Huang 2025) introduced a Fusion-Restoration Model (FRM) that reconstructs multimodal feature representations through a fusion encoder and decoupled decoders, while (Yikang Shi, Xin Zhan, Yaqian Li, Zhongqiang Wu, Wenming Zhang, Haibin Li, 2026) proposed Cycle-CFM, which employs cycle-consistent cross-modal feature mapping and adaptive optimization to improve robustness under noisy industrial environments.

Meanwhile, vision-language models have significantly advanced industrial anomaly detection by enabling zero-shot and few-shot learning. (*Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, Onkar Dabeer*, 2023) demonstrated that CLIP-based representations can achieve competitive zero-shot anomaly classification and

segmentation, whereas (Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, Jiming Chen, 2024) further improved generalization through object-agnostic prompt learning for recognizing anomalies across diverse industrial domains. Beyond visual inspection, lightweight multimodal frameworks (Kang *et al.* 2025) and multimodal digital twin architectures (Ren *et al.* 2026) have further emphasized the importance of efficient multimodal fusion for scalable industrial deployment.

Despite these advances, existing multimodal anomaly detection methods predominantly target visual inspection using RGB images and 3D point clouds. Comparatively little attention has been devoted to heterogeneous cyber–physical operational data, such as PLC telemetry, industrial network traffic, and system logs, which collectively characterize equipment behavior, communication patterns, and control-system events. Furthermore, there remains a limited comparative evaluation of deep multimodal fusion architectures on these heterogeneous industrial modalities, making it difficult to understand how different fusion mechanisms influence anomaly representation and classification performance.

To address this gap, this paper presents a comparative evaluation of four representative multimodal fusion architectures—Fusion MLP, Fusion 1D CNN, Fusion CNN-LSTM, and Cross-Modal Attention Fusion—for industrial anomaly understanding using synchronized PLC telemetry, industrial network traffic, and system log data. Unimodal and classical machine learning baselines are included to assess the contribution of multimodal fusion, while a temporally separated evaluation examines robustness to later observations. In addition to quantitative performance evaluation, latent feature visualization, attention analysis, and modality contribution analysis are employed to investigate how different fusion strategies learn cyber–physical representations. The findings provide practical insights into the role of modality selection and fusion complexity in industrial anomaly understanding and establish a foundation for future context-aware prognostic decision support.

The primary contributions of this paper are summarized as follows:

1. A unified multimodal framework integrating PLC telemetry, industrial network traffic, and system logs for industrial cyber–physical anomaly understanding.

2. A controlled comparative evaluation of four multimodal fusion architectures—Fusion MLP, Fusion 1D CNN, Fusion CNN-LSTM, and Cross-Modal Attention Fusion—together with unimodal and classical machine learning baselines.

3. A temporal robustness evaluation using chronologically separated training, validation, and test partitions to assess performance on later observations.

4. An interpretability analysis using latent feature visualization, cross-modal attention, and modality contribution analysis to investigate learned representations and the relative contribution of individual modalities.

2. METHODOLOGIES

Industrial anomalies can manifest through changes in physical process variables, communication behavior, and control system events. Accordingly, this study evaluates heterogeneous industrial data from PLC telemetry, industrial network traffic, and system logs within a unified experimental framework. As illustrated in Figure 1, the framework consists of synchronized multimodal input, preprocessing, modality feature representation, multimodal fusion, anomaly classification, and performance evaluation. Four multimodal fusion architectures, Fusion MLP, Fusion 1D CNN, Fusion CNN LSTM, and Cross Modal Attention Fusion, are evaluated under a common experimental protocol. Unimodal and classical machine learning baselines are additionally evaluated to determine whether multimodal fusion provides measurable benefits over individual modalities and simpler classifiers.

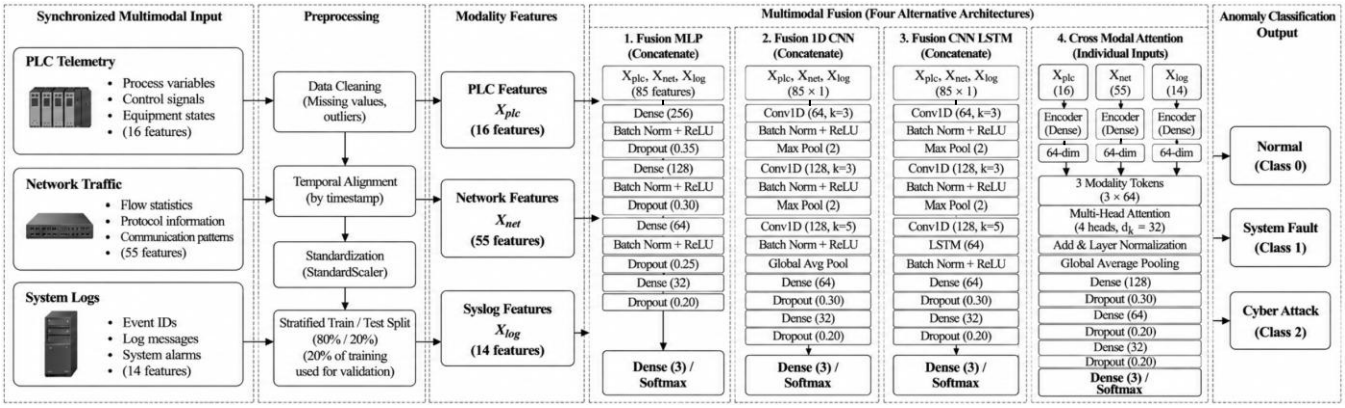


Figure 1. Multimodal industrial anomaly classification framework integrating synchronized PLC telemetry, industrial network traffic, and system logs. The modalities undergo preprocessing before being evaluated using four alternative multimodal fusion architectures for classification into Normal, System Fault, and Cyber Attack classes.

2.1. Proposed Framework

Figure 1 presents the overall multimodal framework and graphical comparison of the four fusion architectures evaluated for industrial cyber-physical anomaly classification. The framework integrates PLC telemetry, industrial network traffic, and system logs, represented by 16, 55, and 14 predictive features, respectively. The modalities are temporally aligned and standardized using parameters estimated exclusively from the training data. The four architectures were selected as representative fusion strategies spanning fully connected, convolutional, recurrent, and attention-based learning. The Fusion MLP concatenates the 85 input features and processes them through fully connected layers containing 256, 128, 64, and 32 units. Fusion 1D CNN operates on the ordered 85-feature representation using four Conv1D layers with 64, 128, 128, and 64 filters, followed by global average pooling and 64 and 32-unit dense layers. Fusion CNN LSTM uses three Conv1D layers with 64, 128, and 128 filters, followed by a 64-unit LSTM and 64 and 32 unit dense layers. In contrast, Cross Modal Attention Fusion maintains separate modality inputs, projects each modality into a 64-dimensional representation through modality-specific encoders, and forms three modality tokens for four-head self-attention with a key dimension of 32. The attended representation is aggregated using global average pooling and processed through 128, 64, and 32-unit dense layers. Each architecture terminates in a three-class softmax classifier for Normal, System Fault, and Cyber Attack classification.

All architectures are evaluated using consistent data partitions, preprocessing procedures, training conditions, and evaluation metrics to support a controlled comparison of the fusion strategies. Unimodal and classical machine learning baselines are additionally evaluated to determine whether increased multimodal fusion complexity provides measurable classification benefits over individual modalities and simpler classifiers.

2.2. Multimodal Data Preparation

The experiments were conducted using the ICS dataset developed by Sudyana *et al.* for anomaly detection in cyber-physical systems. The dataset was generated using an extended Industrial Control System Simulation (ICSSIM) environment that emulates an industrial bottle filling process consisting of two programmable logic controllers (PLCs), two human machine interfaces (HMIs), a water tank, valves, and conveyor belts. Three heterogeneous data sources are considered for multimodal analysis: PLC telemetry, industrial network traffic, and system logs. PLC telemetry captures the physical behavior of the manufacturing process through process variables, sensor measurements, actuator states, and controller signals. Industrial network traffic is provided as packet capture traces and represents communication among industrial devices during system operation. Network flow features are extracted from the packet captures to characterize communication statistics, protocol behavior, and traffic activity. System logs record events generated by the PLCs and HMIs and provide contextual information regarding system activities and operational events.

Because the three modalities are recorded using different timestamp formats and sampling characteristics, all timestamps are converted to a common UTC reference and aligned on a one-second temporal timeline. PLC observations are mapped to the synchronized timeline using exact timestamp matching where available and nearest temporal matching within a maximum tolerance of one second otherwise. Network flows are temporally associated using their flow start timestamps and aggregated within the corresponding one-second intervals. System log events are similarly aggregated according to their timestamps. Intervals

without observed network or system log activity are represented by zero-valued features for the corresponding modality. This procedure produces temporally aligned observations while retaining metadata describing modality availability for auditing purposes; these metadata are excluded from the predictive feature set. Following synchronization and feature screening, the final dataset contains 26,948 observations and 85 predictive features, comprising 16 PLC features, 55 network features, and 14 system log features. The predictive variables were checked for missing values, constant features, duplicate observations, and direct label leakage before model development. The resulting dataset contains no missing predictive values or constant predictive features.

The dataset contains three scenario classes: Normal, System Fault, and Cyber Attack. The final synchronized dataset contains 16,328 Normal observations, 9,505 System Fault observations, and 1,115 Cyber Attack observations, corresponding to approximately 60.59%, 35.27%, and 4.14% of the dataset, respectively. System Fault scenarios include abnormalities such as sensor drift, tank leakage, PLC-related delays, valve malfunction, and memory corruption, while Cyber Attack scenarios include reconnaissance, denial of service, man-in-the-middle, and replay attacks. Before model training, numerical preprocessing parameters are estimated exclusively from the training partition and subsequently applied to the validation and test partitions. Categorical attributes are encoded where applicable. This training-based preprocessing procedure prevents information from the evaluation partitions from influencing feature normalization. The resulting modality-specific representations are denoted as X_{plc}, X_{net}, and $X_{log}$and are used as inputs to the unimodal and multimodal classification experiments. Because network flows are represented according to their flow start timestamps, network activity is sparse relative to the PLC and system log modalities; this representation characteristic is considered when interpreting the unimodal network baseline results.

2.3. Modality Feature Representation

Following multimodal data preparation, each data source is independently transformed into a modality-specific feature representation to preserve its distinct physical, communication, and event information prior to fusion. PLC telemetry is represented by 16 numerical features describing process variables, sensor measurements, actuator states, and controller behavior. Industrial network traffic is represented by 55 features extracted from network flows, including communication statistics, protocol characteristics, traffic volumes, and related flow attributes. System logs are represented by 14 structured features derived from PLC and HMI events, including event source and message-related indicators. The resulting modality representations are denoted as $X_{plc} \in \mathbb{R}^{16}$, $X_{net} \in \mathbb{R}^{55}$, and $X_{log} \in \mathbb{R}^{14}$, yielding a total of 85 predictive features. Identifier fields, absolute timestamps, synchronization metadata, and other nonpredictive attributes are excluded from model inputs. Constant features and features containing missing values are also screened during data preparation to ensure consistent model inputs.

To prevent information leakage, numerical scaling parameters are estimated exclusively from the training partition and subsequently applied without refitting to the validation and test partitions. The three modality-specific feature representations are retained separately for the unimodal experiments and supplied jointly to the multimodal fusion architectures. This common feature representation and preprocessing protocol enables direct comparison of individual modalities and alternative fusion strategies under consistent experimental conditions.

2.4. Multimodal Fusion Architectures

The four architectures were selected as representative fusion strategies spanning fully connected, convolutional, recurrent, and attention-based learning, thereby enabling comparison of increasingly structured approaches to multimodal feature integration. To investigate how different multimodal learning strategies integrate heterogeneous industrial data, four representative fusion architectures are evaluated: Fusion MLP, Fusion 1D CNN, Fusion CNN LSTM, and Cross Modal Attention Fusion. Each architecture uses the same synchronized modality features derived from PLC telemetry, industrial network traffic, and system logs. The same data partitions, preprocessing procedures, training configuration, and evaluation metrics are used to enable a controlled comparison. The architectures differ primarily in how the modality features are integrated and transformed into a fused representation for anomaly classification. In addition to these multimodal models, unimodal and classical machine learning baselines are evaluated separately to determine whether increased fusion complexity provides measurable performance benefits over individual modalities and simpler classifiers.

2.4.1. Fusion MLP

The Fusion MLP performs early fusion by concatenating the PLC, network, and system log feature vectors into a single multimodal representation. The concatenated vector is processed through fully connected layers with nonlinear activation functions to learn relationships among physical process variables, communication characteristics, and system events. Dropout regularization is applied to reduce overfitting, and the learned latent representation is subsequently provided to the classification layer. The Fusion MLP serves as a straightforward neural fusion architecture against which the more structured fusion approaches are compared.

2.4.2. Fusion 1D CNN

The Fusion 1D CNN applies one-dimensional convolutional operations to the concatenated multimodal feature representation. The convolutional filters learn local patterns and interactions within the ordered feature representation, while pooling operations reduce dimensionality and produce higher-level feature maps. The resulting feature maps are aggregated using global average pooling and subsequently processed through fully connected layers to obtain the latent representation used for classification. This architecture evaluates whether convolutional feature extraction provides additional discriminative capability beyond fully connected early fusion

2.4.3. Fusion CNN LSTM

The Fusion CNN LSTM combines convolutional feature extraction with LSTM-based representation learning. The concatenated multimodal features are first processed using convolutional operations to extract local feature patterns. The resulting feature representation is then provided to an LSTM layer to model dependencies within the transformed representation. The LSTM output is subsequently processed through fully connected layers to obtain the latent representation used for anomaly classification. This architecture enables comparison of a combined convolutional and recurrent learning strategy with the MLP and 1D CNN fusion approaches.

2.4.4. Cross-Modal Attention Fusion

The Cross Modal Attention Fusion architecture explicitly models interactions among PLC telemetry, industrial network traffic, and system log representations. Unlike the other fusion architectures, which operate on concatenated features, this model maintains three separate modality inputs. Each modality is independently encoded using a 128-unit fully connected layer with ReLU activation, batch normalization, and 0.30 dropout, followed by a 64-unit fully connected layer and batch normalization. The resulting 64-dimensional representations are reshaped into individual modality tokens and concatenated to form a 3×64multimodal token representation.

Multihead self-attention is then applied across the three modality tokens using four attention heads, a key dimension of 32, and an attention dropout rate of 0.20. A residual connection combines the attention output with the original modality tokens, followed by layer normalization. The resulting representation is aggregated using global average pooling and subsequently processed through fully connected layers containing 128, 64, and 32 units. Batch normalization and dropout are applied within the classification representation, with dropout rates of 0.40, 0.30, and 0.20, respectively. The final 32-dimensional latent representation is passed to a three-class Softmax layer for anomaly classification. The learned attention matrices are additionally analyzed to characterize how information is allocated among the PLC, network, and system log representations. These attention weights are interpreted as indicators of the model's internal information routing rather than as causal measures of modality importance.

2.5. Anomaly Classification

The latent representation produced by each multimodal fusion architecture is passed to a fully connected classification layer with a Softmax activation function. For each input observation, the Softmax layer estimates the probability associated with each class, and the class with the highest predicted probability is selected as the final classification. The supervised classification task distinguishes three scenario classes: Normal, System Fault, and Cyber Attack. Normal represents nominal operation of the industrial process under fault-free conditions. System Fault represents abnormalities associated with equipment or process faults, while Cyber Attack represents observations collected under cybersecurity attack scenarios. The same three-class definition is maintained across the multimodal, unimodal, and classical supervised experiments to support consistent comparison of classification performance.

2.6. Baseline Models

To assess whether multimodal fusion provides additional classification benefit over individual data sources, three unimodal neural baselines are evaluated using PLC telemetry, industrial network traffic, and system logs independently. The PLC Only, Network Only, and Syslog Only models use the same corresponding modality features and experimental partitions as the multimodal models, enabling direct assessment of the discriminative information available within each modality. Logistic Regression and Decision Tree classifiers are additionally evaluated using the complete predictive feature set as classical supervised baselines. These models provide lower complexity reference points for determining whether classification performance is attributable to sophisticated fusion architectures or to inherent separability within the dataset. An Isolation Forest model trained using Normal observations is also evaluated as an unsupervised novelty detection baseline. Because Isolation Forest performs binary Normal versus Anomaly detection rather than three-class classification, its performance is analyzed separately from the supervised classifiers.

2.7. Model Training and Experimental Protocol

All neural models were trained using the Adam optimizer with an initial learning rate of 0.001, a batch size of 128, and a maximum of 100 epochs. Categorical cross-entropy was used as the loss function, and balanced class weights were computed from the training data to account for class imbalance. Numerical scaling parameters were estimated

exclusively from the training partition and subsequently applied to the validation and test partitions to prevent information leakage. For the primary evaluation, the synchronized dataset was divided using an 80/20 stratified train-test split with a fixed random seed of 42. Twenty percent of the training partition was used for validation during neural model training. The same data partitioning protocol was maintained across the unimodal and multimodal neural models to enable consistent comparison. Early stopping was applied based on validation loss with a patience of 12 epochs, and the model parameters corresponding to the lowest validation loss were restored before test evaluation. The learning rate was reduced by a factor of 0.5 after five consecutive epochs without improvement, with a minimum learning rate of 10^{-6}.

To examine whether the high classification performance obtained under random stratification persisted when temporal separation was imposed, an additional chronological robustness evaluation was conducted. For each scenario independently, observations were ordered chronologically and partitioned into the earliest 64% for training, the subsequent 16% for validation, and the final 20% for testing. No observations were randomly exchanged across these temporal partitions. Numerical scaling parameters and class weights were estimated exclusively from the temporal training partition and applied to the corresponding validation and test data. The same neural architectures and training configuration used in the primary evaluation were maintained for the temporal experiment. The temporal evaluation assesses generalization to later observations within each recorded scenario and reduces the influence of temporal proximity between training and test samples. However, because the dataset contains a limited number of independent scenario runs, this experiment should be interpreted as within-scenario temporal robustness rather than cross-run or cross-domain generalization.

2.8. Evaluation Metrics

Classification performance was evaluated using accuracy, macro precision, macro recall, macro F1 score, weighted F1 score, macro ROC AUC, and weighted ROC AUC. Because the synchronized dataset is class-imbalanced, macro-averaged metrics were included to assign equal importance to Normal, System Fault, and Cyber Attack classes, while weighted metrics account for the relative number of observations within each class. Confusion matrices were additionally generated to examine class-specific prediction errors.

For a multiclass problem with C classes, precision, recall, and F1 score were first computed independently for each class using a one versus rest formulation. For class c,

$$P_c = \frac{TP_c}{TP_c + FP_c} \tag{1}$$

$$R_c = \frac{TP_c}{TP_c + FN_c} \tag{2}$$

and

$$F1_c = \frac{2P_cR_c}{P_c + R_c} \tag{3}$$

where TP_c, FP_c, and FN_c denote the true positive, false positive, and false negative counts for class c, respectively. In general, accuracy is defined as:

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}.$$

Macro-averaged metrics were calculated as the unweighted mean across the C classes. For example,

$$Macro\ F1 = \frac{1}{C}\sum_{c=1}^{C} F\,1_c \tag{4}$$

Weighted F1 was calculated according to the support n_c of each class:

$$Weighted\ F1 = \sum_{c=1}^{C} \frac{n_c}{N} F1_c \tag{5}$$

where N is the total number of test observations. ROC AUC was calculated using a one-versus-rest formulation from the predicted class probabilities, with both macro and class support weighted averages reported.

The same supervised evaluation metrics were applied to the PLC Only, Network Only, Syslog Only, Logistic Regression, Decision Tree, Fusion MLP, Fusion 1D CNN, Fusion CNN LSTM, and Cross Modal Attention Fusion models to enable direct comparison. The Isolation Forest baseline was evaluated separately as a binary Normal versus Anomaly detector using accuracy, anomaly precision, anomaly recall, anomaly F1 score, ROC AUC, and a binary confusion matrix because its task formulation differs from the three-class supervised classification problem. For the four multimodal neural architectures, the learned latent representations were additionally examined using t-distributed stochastic neighbor embedding (t-SNE). A 32-dimensional latent representation preceding the final classification stage was extracted from each model and standardized before projection into two dimensions. The same t-SNE configuration was used across all four models, with a perplexity of 30, PCA initialization, and a fixed random seed of 42. The visualization was used as a qualitative analysis of the organization of the learned representation space rather than as an independent measure of classification performance.

For the Cross Modal Attention Fusion model, the learned attention matrices were further analyzed to characterize interactions among the PLC, network, and system log

representations. Mean attention allocation was examined across modalities and classes. These attention values were interpreted as indicators of the model's internal information routing and not as causal measures of modality importance.

3. EXPERIMENTS

A comprehensive experimental study was conducted to evaluate multimodal anomaly classification using synchronized PLC telemetry, industrial network traffic, and system logs. The evaluation considered three groups of supervised models: unimodal neural baselines using PLC, network, and system log features independently; classical machine learning baselines consisting of Logistic Regression and Decision Tree; and four multimodal architectures comprising Fusion MLP, Fusion 1D CNN, Fusion CNN LSTM, and Cross Modal Attention Fusion. An Isolation Forest model trained on normal observations was additionally evaluated as an unsupervised binary anomaly detection baseline.

The experiments were designed to address two primary questions. First, the unimodal and classical baselines were used to determine whether multimodal fusion provides measurable classification benefits over individual modalities and lower complexity models. Second, the four fusion architectures were compared to examine whether increasing architectural complexity improves multimodal classification performance. In addition to the primary stratified evaluation, a temporally separated robustness analysis was conducted to determine whether the observed classification performance persisted when models were evaluated on later observations within each scenario. The detailed data partitioning, model training, and evaluation procedures are described in Section 2.7.

3.1. Experimental Setup

The experiments were conducted on a Linux workstation running Ubuntu 22.04.5 LTS (64-bit). The workstation was equipped with an AMD Ryzen Threadripper Pro 5955WX processor with 16 cores and 32 threads, 128 GB of system memory, and an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory. NVIDIA Driver version 570.211.01 and CUDA 12.8 were used to support GPU-accelerated model training. The neural models were implemented in Python 3.10.12 using TensorFlow 2.21.0. All models within each comparative experiment were executed under the same hardware and software environment. Consistent data partitions and preprocessing procedures were maintained across comparable models as described in Section 2.7, enabling differences in predictive performance to be evaluated under controlled experimental conditions. To support reproducibility, the source code and experimental implementation are publicly available in the accompanying GitHub (https://github.com/Lukmon892/multimodal-ics-anomaly-detection/tree/main)

4. RESULTS

The experimental evaluation compared unimodal, classical machine learning, and multimodal fusion approaches for three-class industrial anomaly classification using synchronized PLC telemetry, industrial network traffic, and system logs. The primary evaluation used the stratified partition described in Section 2.7, followed by a temporally separated robustness evaluation to assess whether the observed performance persisted for later observations within each scenario. Performance was assessed using accuracy, macro precision, macro recall, macro F1 score, weighted F1 score, and ROC AUC. Confusion matrices were used to examine class-specific errors, while ROC curves and t-SNE visualizations were used to assess class discrimination and the organization of learned latent representations. For the Cross Modal Attention Fusion model, attention analysis was additionally performed to examine the allocation of attention among the three modality representations. The following subsections first evaluate the unimodal and classical baselines, followed by the multimodal fusion architectures and temporal robustness analysis.

4.1. Unimodal and Classical Baseline Performance

Table 1 compares the unimodal, classical, and multimodal models under the stratified evaluation. The results show substantial differences in the discriminative capability of the individual modalities. The PLC Only model achieved 98.85% accuracy and a macro F1 score of 95.48%, while the Syslog Only model achieved 99.93% accuracy and 99.74% macro F1. In contrast, the Network Only model achieved 60.61% accuracy and 30.20% macro F1, indicating substantially weaker class discrimination under the evaluated network representation. The classical baselines also achieved near-perfect performance. Logistic Regression obtained 99.91% accuracy and 99.66% macro F1, while the Decision Tree achieved 99.93% accuracy and 99.74% macro F1. These results demonstrate that the benchmark is highly separable under the stratified evaluation and that high classification performance is not limited to deep multimodal architectures.

Importantly, the unimodal results show that multimodal fusion is not required to achieve high aggregate classification performance on this dataset. In particular, the Syslog Only model performed comparably to the multimodal architectures. Consequently, the near-perfect performance of the fusion models should not be interpreted solely as evidence of superior multimodal learning. Instead, the results indicate that individual modalities, particularly system logs, contain highly discriminative information for the evaluated scenarios.

4.1.1. Unsupervised Anomaly Detection Baseline

Isolation Forest, trained using only Normal observations, achieved 93.25% binary accuracy, 85.37% anomaly precision, 100.00% anomaly recall, 92.11% anomaly F1, and

98.87% ROC AUC. The model detected all fault and cyber-attack observations when treated collectively as anomalies but generated 364 false positives among Normal observations. These results demonstrate that unsupervised novelty detection can identify deviations from nominal behavior without labeled anomaly classes, although with substantially more false alarms than the supervised classifiers.

4.2. Multimodal Fusion Performance

All four multimodal architectures achieved high classification performance under the stratified evaluation. Fusion MLP achieved 99.81% accuracy and 99.32% macro F1, while Fusion CNN LSTM achieved 99.94% accuracy and 99.75% macro F1. Cross Modal Attention Fusion achieved 99.96% accuracy and 99.84% macro F1. Fusion 1D CNN obtained perfect classification on the stratified test partition.

Although the fusion architectures achieved consistently high performance, their improvements relative to the strongest unimodal and classical baselines were small. For example, Syslog Only and the Decision Tree each achieved 99.93% accuracy and 99.74% macro F1. Therefore, increasing fusion architecture complexity did not produce a consistent improvement in aggregate classification performance. Rather, the results indicate that the synchronized multimodal representation supports accurate classification while the benchmark itself contains strong discriminative signals that can also be exploited by individual modalities and simpler classifiers.

Table 1. Performance comparison of unimodal, classical, and multimodal models under the stratified evaluation

Model	Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro F1 (%)	Weighted F1 (%)	Macro ROC AUC
PLC Only	98.85	93.23	98.25	95.48	98.90	99.96
Network Only	60.61	33.01	36.27	30.20	46.36	51.74
Syslog Only	99.93	99.54	99.94	99.74	99.93	99.99
Logistic Regression	99.91	99.79	99.52	99.66	99.91	99.81
Decision Tree	99.93	99.54	99.94	99.74	99.93	100.00
Fusion MLP	99.81	99.19	99.45	99.32	99.81	99.95
Fusion 1D CNN	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Fusion CNN LSTM	99.94	99.56	99.95	99.75	99.94	99.99
Cross Modal Attention	99.96	99.70	99.97	99.84	99.96	100.00

Figure 2 presents the confusion matrices for the four multimodal fusion architectures under the stratified test evaluation. All models achieved strong class-specific performance across the Normal, System Fault, and Cyber Attack classes. As shown in Fig. 2(a), Fusion MLP correctly classified 3,263 of 3,266 Normal, 1,897 of 1,901 System Fault, and 220 of 223 Cyber Attack observations, resulting in ten misclassifications. Fusion 1D CNN, shown in Fig. 2(b), correctly classified all 5,390 test observations, with no errors across the three classes. Fusion CNN LSTM in Fig. 2(c) produced only three errors, all corresponding to System Fault observations classified as Cyber Attack. Cross Modal Attention Fusion in Fig. 2(d) produced only two errors, with one Normal observation and one System Fault observation incorrectly classified as Cyber Attack.

Collectively, the confusion matrices confirm that the four fusion architectures provide strong discrimination among the three scenario classes under the stratified evaluation. However, the similarly high performance observed for the unimodal and classical baselines indicates that these low error rates should not be attributed solely to multimodal fusion. Rather, they further demonstrate the strong class separability present in the evaluated dataset.

Table 2. Performance under temporally separated 64/16/20 evaluation

Model	Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro F1 (%)	Weighted F1 (%)	Macro ROC AUC
Logistic Regression	99.81	98.57	99.82	99.18	99.82	100.00
Decision Tree	99.83	98.71	99.84	99.26	99.83	99.89
PLC Only	**99.93**	**99.96**	99.40	**99.68**	**99.93**	**100.00**
Network Only	35.53	48.94	36.43	22.98	19.33	51.90
Syslog Only	99.81	98.57	99.82	99.18	99.82	99.82
Fusion MLP	99.80	98.56	99.81	99.17	99.80	99.98
Fusion 1D CNN	99.81	98.57	99.82	99.18	99.82	99.92
Fusion CNN LSTM	99.81	98.57	99.82	99.18	99.82	99.80
Cross-Modal Attention	99.78	98.30	99.80	99.03	99.78	99.92

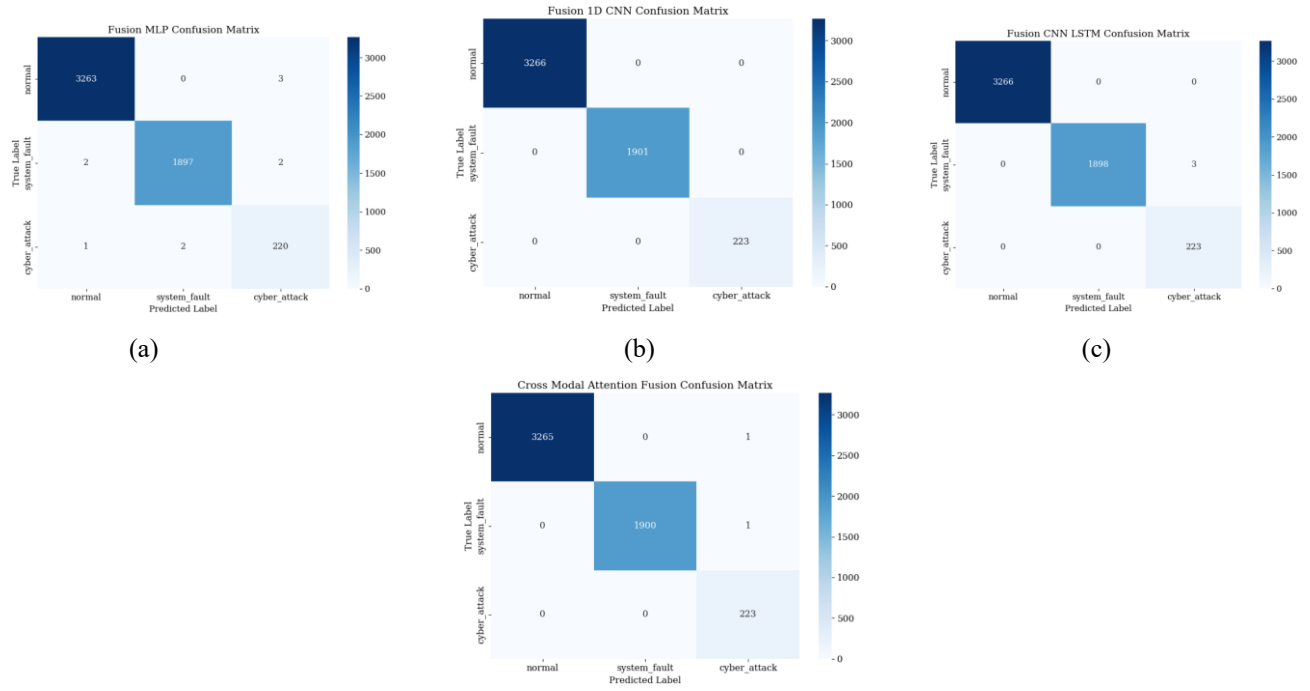
Table 2 presents the results of the temporally separated evaluation, in which the earliest 64% of each scenario was used for training, the subsequent 16% for validation, and the final 20% for testing. High classification performance largely persisted under temporal separation, indicating that the separability observed under random stratification cannot be explained solely by temporal adjacency between randomly partitioned observations. However, the temporal experiment further demonstrates that multimodal fusion does not consistently outperform the individual modalities. The PLC Only model achieved the strongest temporal performance, with 99.93% accuracy and 99.68% macro F1, exceeding the four fusion architectures, whose macro F1 scores ranged from 99.03% to 99.18%. Syslog Only, Logistic Regression, and Decision Tree also maintained approximately 99.8% accuracy. In contrast, Network Only performance decreased to 35.53% accuracy and 22.98% macro F1. This result indicates that the network modality, under the current sparse flow start-based temporal representation, provides substantially weaker standalone discrimination than PLC telemetry and system logs.

Overall, the temporal evaluation confirms that the benchmark contains highly persistent discriminative signatures. The results therefore do not support a claim that increasing multimodal fusion complexity necessarily improves aggregate classification performance. Instead, they demonstrate the importance of evaluating individual modalities and simpler baselines when assessing multimodal industrial anomaly classification systems.

Figure 3 presents the t-SNE projections of the 32-dimensional latent representations learned by the four multimodal fusion architectures. Across all four models, Normal, System Fault, and Cyber Attack observations occupy largely distinct regions of the projected feature space. In particular, Cyber Attack observations form comparatively compact and well-separated clusters, while Normal and System Fault observations are also substantially separated despite greater within-class variation. These patterns are consistent with the high classification performance and low misclassification rates observed in the quantitative results. However, the clear class separation should be interpreted together with the unimodal and classical baseline results. Since individual modalities and simpler classifiers also achieved high classification performance, the t-SNE results suggest that the evaluated scenarios contain strongly discriminative features that are preserved by the learned fusion representations. Thus, the visual separation supports the observed classification results but does not by itself demonstrate that increased multimodal fusion complexity is necessary for class discrimination.

Figure 4 illustrates the attention allocation learned by the Cross Modal Attention Fusion model. As shown in Fig. 4(a), PLC and Syslog queries allocate greater attention to the Syslog representation, whereas the Network query distributes attention relatively evenly across the three modalities. Overall, Fig. 4(b) shows that Syslog receives the highest

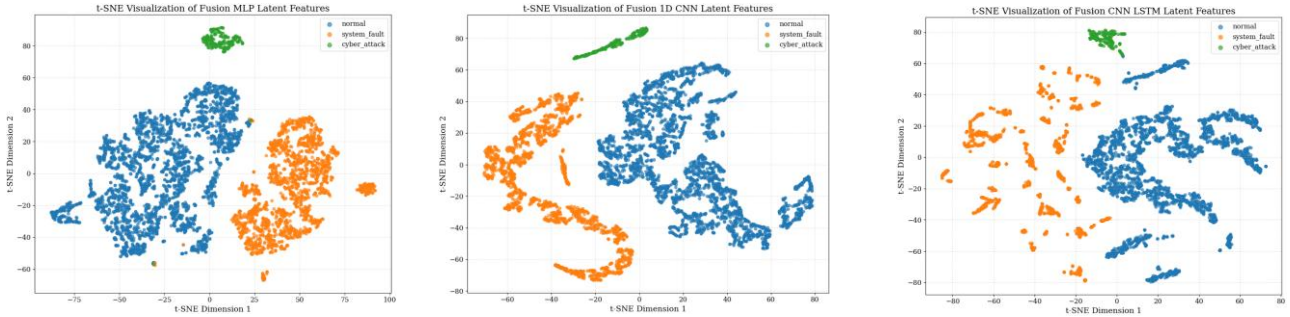
mean attention (0.3730), followed by PLC (0.3418) and Network (0.2852). This pattern is broadly consistent with the unimodal evaluation, in which Syslog provides substantially stronger standalone classification performance than Network under the stratified evaluation. The results indicate that the attention model allocates information unequally across the available modality representations. However, the attention weights are interpreted as indicators of internal information routing rather than causal measures of modality importance.



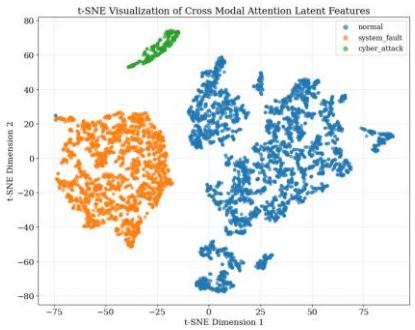
(a) (b) (c)

(d)

Figure 2. Confusion matrices for (a) Fusion MLP, (b) Fusion 1D CNN, (c) Fusion CNN LSTM, and (d) Cross Modal Attention Fusion under the stratified test evaluation. All architectures demonstrate strong class-specific performance, with Fusion 1D CNN correctly classifying all test observations.

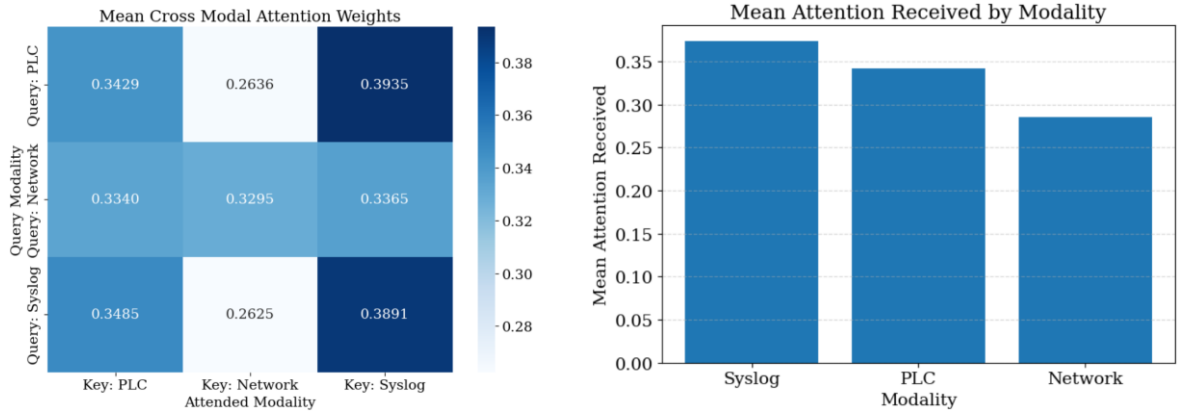


(a) (b) (c)



(d)

Figure 3. t-SNE visualization of the learned 32-dimensional latent feature representations for (a) Fusion MLP, (b) Fusion 1D CNN, (c) Fusion CNN LSTM, and (d) Cross-Modal Attention Fusion.



(a) Cross-modal attention weights (b) Mean attention received by each modality

Figure 4. Cross-modal attention analysis showing (a) mean attention weights among PLC, Network, and Syslog representations and (b) mean attention received by each modality. Syslog receives the highest average attention, followed by PLC and Network, consistent with differences observed among the unimodal baselines

5. DISCUSSION

The results demonstrate high classification performance across the four multimodal fusion architectures; however, the unimodal and classical baselines indicate that this performance cannot be attributed solely to multimodal fusion. Under the stratified evaluation, Syslog Only achieved 99.93% accuracy and 99.74% macro F1, while Logistic Regression and Decision Tree achieved 99.91% and 99.93% accuracy, respectively. These results indicate strong class separability within the ICSSIM benchmark and show that increased fusion complexity provides limited improvement when individual modalities contain highly discriminative signals. The temporal robustness analysis produced a similar conclusion. Using chronologically separated 64% training, 16% validation, and 20% testing partitions, PLC Only achieved the strongest performance with 99.93% accuracy and 99.68% macro F1, whereas the four fusion architectures achieved macro F1 scores between 99.03% and 99.18%. In contrast, Network Only achieved 35.53% accuracy and 22.98% macro F1, indicating substantially weaker standalone discrimination under the current sparse flow-based network representation. The persistence of high performance under temporal separation suggests that random temporal adjacency alone does not explain the strong classification results.

The confusion matrices and t-SNE projections further demonstrate low misclassification rates and clear class separation in the learned latent representations. Cross-modal attention analysis showed that Syslog received the highest mean attention (0.3730), followed by PLC (0.3418) and Network (0.2852), broadly consistent with the unimodal results under the stratified evaluation. These attention weights characterize the model's internal information routing and are not interpreted as causal measures of modality importance.

Collectively, the findings show that multimodal fusion effectively integrates heterogeneous industrial data, but

greater architectural complexity does not necessarily improve aggregate classification performance on this benchmark. The results emphasize the importance of unimodal baselines, simpler classifiers, and temporal robustness evaluation when assessing multimodal industrial anomaly detection systems. Since the temporal experiment evaluates later observations within the same recorded scenarios, future work will investigate independent operating runs, more diverse industrial datasets, improved network representations, and conditions involving noisy, incomplete, or less discriminative modalities.

6. CONCLUSION

This paper presented a controlled evaluation of four multimodal fusion architectures for industrial anomaly classification using synchronized PLC telemetry, industrial network traffic, and system logs. Unimodal and classical baselines, together with temporal robustness evaluation, were incorporated to assess the contribution of multimodal fusion beyond individual modalities and simpler classifiers. The results show that although multimodal fusion supports accurate integration of heterogeneous industrial data, increased architectural complexity does not necessarily provide additional classification benefit when individual modalities contain highly discriminative signals. The study therefore highlights the importance of evaluating multimodal models against strong unimodal and classical baselines rather than attributing high performance solely to fusion complexity. Future work will extend the evaluation to independent operating runs, more diverse industrial environments, improved network representations, and conditions involving noisy, incomplete, or less discriminative modalities.

REFERENCES

Ren, C., Li, M., Chen, C., Guan, X., & Huang, G. Q. (2026). Multi-modal digital twins for industrial anomaly detection: Framework, method, and application. *Robotics and Computer-Integrated Manufacturing*, *97*, 103068.

Kang, J., Lee, J., & Jeong, J. (2025). KD-LightMAD: Knowledge Distillation–based Lightweight Multimodal Anomaly Detection Framework for Smart Manufacturing. *IEEE Access*.

Wang, Y., Peng, J., Zhang, J., Yi, R., Wang, Y., & Wang, C. (2023). Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8032-8041).

Cheng, H., Luo, J., & Zhang, X. (2025). Multimodal industrial anomaly detection via uni-modal and cross-modal fusion. *IEEE Transactions on Industrial Informatics*.

Wang, J., Niu, Y., & Huang, B. (2025). Fusion-restoration model for industrial multimodal anomaly detection. *Neurocomputing*, *637*, 130073.

Shi, Y., Zhan, X., Li, Y., Wu, Z., Zhang, W., & Li, H. (2025). Cycle-CFM: An Unsupervised Framework for Robust Multimodal Anomaly Detection in Industrial Settings. *Expert Systems with Applications*, 129745.

Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., & Dabeer, O. (2023). Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 19606-19616).

Zhou, Q., Pang, G., Tian, Y., He, S., & Chen, J. (2024, May). Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations* (Vol. 2024, pp. 49705-49737).

Sudyana, D. (2025). ICS Dataset [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.17165850