International Journal on Advanced Science Engineering Information Technology

Vol.15 (2025) No. 1
ISSN: 2088-5334

Establishment of a Real-Time Risk Assessment and Preventive Safety Management System in Industrial Environments Utilizing Multimodal Data and Advanced Deep Reinforcement Learning Techniques

Hyun Sim [a,*], Hyunwook Kim [b]

[a] *Department of Smart Agriculture, Sunchon National University, Republic of Korea*
[b] *Kornerstone. Co,Ltd., Suncheon City, Republic of Korea*
Corresponding author: [*]simhyun@scnu.ac.kr

***Abstract*—This study proposes a new paradigm for real-time, predictive, and multidimensional risk assessment in industrial environments by leveraging multimodal data (video, audio, environmental sensors). Existing risk assessment systems typically rely on single data sources or subjective judgment, making it difficult to adapt swiftly to complex workplace changes and achieve real-time responsiveness. To address these limitations, we constructed a large-scale multimodal dataset of approximately 10TB, employing real-time streaming-based preprocessing and data synchronization. This approach integrates various data sources—high-resolution cameras, high-sensitivity microphones, and environmental sensors (temperature, humidity, vibration)—and applies data augmentation techniques such as AutoAugment and MixUp to build robust models capable of handling diverse environmental conditions. We adopted a hybrid analytical algorithm combining Vision Transformer (ViT) and YOLOv8, achieving high accuracy (over 95%) and real-time processing (average response time under one second). Additionally, we utilized machine learning algorithms such as SVM, Random Forest, and K-Means to detect anomalies in audio and environmental sensor data, thus identifying latent risk factors. Experimental results demonstrate multifaceted performance improvements compared to conventional approaches, including over a 15% increase in accuracy, approximately 30% reduction in response time, about 20% reduction in power consumption, and user satisfaction exceeding 90%. These achievements were verified across various industrial settings—chemical, manufacturing, logistics—highlighting the system's capacity to detect complex risk factors and respond proactively. By seamlessly integrating multimodal data analysis, state-of-the-art deep learning models (ViT, YOLOv8), and reinforcement learning-based response strategies, we have demonstrated a transition from traditional, static, and retrospective risk assessment to an intelligent, real-time, and predictive safety management framework.**

***Keywords*— Real-time risk assessment; preventive safety management; multimodal data; YOLOv8; model lightweighting.**

Manuscript received 11 Oct. 2024; revised 8 Dec. 2024; accepted 12 Jan. 2025. Date of publication 28 Feb. 2025.

I. INTRODUCTION

Industrial sites are complex environments where various hazards coexist, such as heavy machinery operation, high-temperature/high-pressure equipment, handling of hazardous materials, and intricate logistics flows. Accidents in these settings can lead to severe human casualties and economic losses. Accordingly, international standards like ISO 45001 and regional occupational safety and health regulations (e.g., OSHA guidelines) emphasize systematic and quantitative risk assessments for workplace safety. Traditional risk assessment systems rely heavily on evaluators' experience, partial observations, or simple rule-based/statistical models, making it challenging to swiftly adapt to the irregularities and changes in real-world work environments and comprehensively detect complex risk factors. For example, suppose an existing system attains only about 80% hazard detection accuracy or issues alerts several seconds after an abnormal situation occurs. In that case, its practical contribution to accident prevention may be limited.

Rapid advancements in deep learning and the Internet of Things (IoT) have recently offered new paradigms to overcome these limitations. Deep learning automatically learns complex patterns from large-scale data, achieving high recognition performance. In particular, object recognition and inference speed improvements have created favorable conditions for real-time hazard detection. State-of-the-art algorithms such as Vision Transformer (ViT) and YOLOv8 achieve higher accuracy and faster processing times than traditional CNN-based models (e.g., Faster R-CNN), and

ongoing efforts toward model lightweight aim to enhance energy efficiency. Meanwhile, the widespread adoption of IoT-based sensor networks simplifies the collection of various forms of multimodal data—video, audio, temperature, humidity, and vibration—thereby breaking the dependency on a single data source and offering opportunities to capture complex risk situations more precisely. However, most existing deep learning-based risk assessment studies focus on a single modality (e.g., images) or improving a specific model's performance, with relatively few investigations into the complementary use of diverse data. Moreover, research that comprehensively considers practical constraints—such as real-time responsiveness, energy efficiency, and equipment availability demanded by actual industrial fields—is still limited.

This study aims to radically improve the precision and real-time capability of conventional risk assessment systems through the integrated use of multimodal data (video, audio, environmental sensor data) combined with state-of-the-art deep learning algorithms (ViT, YOLOv8), ultimately proposing a new standard applicable to various industrial settings. More specifically, we set the following goals:

1) Improved Accuracy: By leveraging multimodal data and ensemble analysis of the latest models, raise hazard detection accuracy to over 95%, achieving more than a 15% improvement compared to existing approaches.

2) Enhanced Real-Time Response: Apply model lightweight and IoT-based real-time streaming technologies to maintain an average response time below one second, enabling rapid response strategies.

3) Automated Contextual Response: Utilize reinforcement learning-based decision-making models to automatically generate optimal response strategies for detected hazards under various environmental conditions and provide intuitive alerts through user interfaces.

Through this study, we aim to improve hazard detection capabilities and make a practical contribution to ensuring worker safety. Ultimately, the proposed system can enhance overall safety management efficiency and prevent accidents across the industrial sector.

This paper is composed of four main sections. Section I (Introduction) presents the research background, necessity, objectives, and overall motivation for this study. Section II (Materials and Method) reviews the limitations of existing risk assessment technologies, recent deep learning algorithms (Vision Transformer, YOLOv8), and the potential of multimodal data. It also details the multimodal data collection/processing strategies, deep learning model architectures optimization, and reinforcement learning-based response plan proposals. Section III (Results and Discussion) describes the integrated hardware/software environment, dataset construction, real-world industrial application cases, and results and provides a comprehensive performance evaluation (accuracy, processing speed, resource efficiency, user satisfaction). It also summarizes the research findings, discusses the effects of multimodal analysis and advanced deep learning applications, and suggests limitations and future directions. Section IV (Conclusion) concludes the study by summarizing its significance and evaluating its potential contributions and applications in industrial safety.

II. Materials and Method

Using this document as a camera-ready template is an easy way to comply with the conference paper formatting requirements.

A. Trends in Risk Assessment Technology for Industrial Safety

Industrial accidents in the workplace can lead to human casualties and economic losses, prompting international standards (e.g., ISO 45001) and regional occupational safety and health regulations (e.g., OSHA) to emphasize systematic, quantitative risk assessments and preventive measures. Traditional approaches have relied on expert knowledge, checklists, and statistical analyses, effectively identifying and mitigating potential hazards preemptively. However, such methods struggle to adapt to dynamic, complex environments and real-time response needs, often relying on subjective judgment that may vary between practitioners.

Recent advancements in deep learning have catalyzed the development of more robust and automated risk assessment tools. Anomaly detection and out-of-distribution (OOD) detection techniques, studied initially for industrial defect detection or medical diagnostics [1], [4], [12], [24], [25], [28],[29], [30] have shown significant potential in enhancing safety monitoring. Additionally, self-supervised approaches [7], [10], [19], [21], and generative methods [20], [30] are being investigated to improve model robustness against unseen or changing industrial conditions. The increasing ubiquity of IoT devices and real-time data pipelines further enables large-scale monitoring and hazard prevention, integrating sensor data for anomaly detection or OOD detection in ever-evolving industrial contexts [6], [16], [18], [23].

B. Object Detection Algorithms and Deep Learning-Based Hazard Recognition

Object detection is essential for locating and classifying workers, machinery, and moving vehicles in industrial facilities. Traditional two-stage detectors like Faster R-CNN have achieved high accuracy but often lack the real-time performance needed in dynamic industrial settings. Recently, the YOLO (You Only Look Once) series has gained traction due to its balance between speed and accuracy [24], [25]. State-of-the-art implementations (e.g., YOLOv8) focus on further optimizing computational efficiency to accommodate on-edge or resource-constrained environments.

In parallel, transformer-based models have opened new horizons in computer vision. Vision Transformer (ViT) [5], processes images as a sequence of patches, enabling global context modeling that can outperform conventional convolutional neural networks in classification and detection tasks. These methods better capture complex work environments with multiple overlapping risk factors by leveraging attention mechanisms. Beyond ViT, approaches like Detection Transformer (DETR) reframe detection tasks as a direct set prediction problem, simplifying the detection pipeline. Coupled with large-scale pretraining or transfer learning approaches such as Big Transfer (BiT) [9], these

methods can further improve accuracy and robustness in industrial scenarios.

C. Advancing Risk Assessment through Multimodal Data Utilization

While visual data have been a primary focus in industrial risk assessment, relying solely on images can be insufficient for comprehensive hazard detection, particularly in low-visibility conditions or scenarios involving abnormal sounds and vibrations. Recent works integrate environmental sensing with audio inputs and positional data to bolster anomaly detection capabilities [2], [3], [13], [22]. By merging heterogeneous signals—temperature, humidity, vibration, or acoustic signatures—these systems can detect subtle anomalies (e.g., incipient machinery failure or unsafe working conditions) that are not visually observable [1], [14], [29].

Self-supervised or contrastive learning paradigms can further enhance multimodal feature representations, especially when labeled data are scarce [2], [8], [21]. Methods like CutPaste [10] or mean-shifted contrastive loss [15] demonstrate how unsupervised or minimally supervised data augmentations can capture rich feature distributions, leading to more accurate anomaly or OOD detection. Moreover, combining audio-visual cues with reinforcement learning-based controllers has been proposed to generate proactive recommendations or interventions, reducing response latency in critical situations [11], [17].

D. Limitations of Previous Studies and the Distinctiveness of This Research

Despite these advances, many studies still focus on single-modal inputs or lack real-time responsiveness. Conventional anomaly detection methods may be ill-suited to large-scale, continuously changing industrial environments without dedicated strategies for domain adaptation or fine-tuning [16], [19]. Overconfidence of neural networks in out-of-distribution settings also remains challenging, prompting techniques like logit normalization [23] or likelihood ratio methods [20] to calibrate uncertainty estimations.

Our work integrates multimodal data—including video, audio, and various environmental sensor streams—to overcome these limitations within a unified risk assessment framework. We adopt recent deep learning algorithms (ViT [5], YOLOv8 [24], [25]) and advanced anomaly/OOD detection strategies [1], [6], [10], [15] to achieve both high accuracy and real-time performance. We propose a novel approach that moves beyond reactive detection to comprehensive predictive risk management by leveraging reinforcement learning to recommend proactive response measures.

E. Overview of the System Architecture

The proposed system consists of four main modules: (1) Multimodal Data Collection Module, (2) Data Processing and Training Module, (3) Deep Learning-Based Analysis Module, and (4) Result Visualization and Alert Module. Through the organic linkage of these modules, the system acquires high-quality data in real time. It applies state-of-the-art deep learning algorithms and reinforcement learning to reliably manage complex risk factors in industrial fields.

Multimodal Data Collection Module: High-resolution cameras, high-sensitivity microphones, and IoT sensors (temperature, humidity, vibration) are employed to collect video, audio, and environmental data in a real-time streaming format. Such multi-source data acquisition enables stable hazard detection even under visual limitations (poor lighting, occlusion) and facilitates the early identification of potential risks (e.g., equipment overheating and vibration anomalies).

Data Processing and Training Module: The collected data are transmitted and normalized efficiently using distributed processing frameworks such as Apache Kafka and Spark. Data augmentation techniques like AutoAugment and MixUp are applied to reflect diverse environmental changes, while temporal synchronization and outlier removal ensure training stability. This approach maximizes the diversity and reliability of the training dataset.

Deep Learning-Based Analysis Module: A hybrid model structure combining the Vision Transformer (ViT) and YOLOv8 is adopted. ViT ensures high accuracy by capturing global contextual features, and YOLOv8 guarantees real-time detection with its lightweight architecture. Subsequently, SVM is used to detect anomalous audio signals, and Random Forest or K-Means is employed to detect environmental sensor data anomaly. An ensemble technique integrates the inference results from each modality to produce a comprehensive risk profile.

Result Visualization and Alert Module: The analysis results are presented through an intuitive user interface, and a real-time alert system enables immediate responses from operators and managers. Beyond listing detected hazards, this module uses a reinforcement learning-based decision-making model to automatically propose optimal response strategies, thereby implementing a preventive safety management framework.

F. Data Collection and Processing

Over approximately six months, multimodal data were collected from various industrial environments (manufacturing, logistics, construction), totaling about 7.5TB (video: 3TB, audio: 2TB, environmental sensors: 2.5TB).

TABLE I
TYPES AND CHARACTERISTICS OF COLLECTED MULTIMODAL DATA

Data Type	Collection Equipment	Key Features
Video Data	High-resolution camera	Real-time video, various angles
Audio Data	High-sensitivity microphone	Includes mechanical noises, abnormal sound patterns
Environmental Data	Temperature, humidity, vibration sensors	Real-time environmental change detection

TABLE II
DATASET COMPOSITION AND SCALE

Industry	Data Size	Main Risk Factors
Manufacturing	3TB	Machine malfunction, worker carelessness
Logistics	2TB	Vehicle collisions, falling cargo
Construction	2.5TB	Falls, falling objects, equipment failures

As shown in [Table II] and [Table IIIIV], the data collected via real-time streaming are delivered to the collection module through Kafka queues and processed through a Spark-based preprocessing pipeline. This pipeline removes noise, normalizes the data, and synchronizes timestamps, correcting temporal discrepancies among data sources and minimizing external disturbances. Consequently, the model can learn patterns stably. Moreover, data augmentation techniques such as AutoAugment and MixUp create a training environment robust to varying lighting conditions, noise levels, and background changes.

G. Deep Learning Model Development and Optimization

This study employs Vision Transformer (ViT) and YOLOv8 as core analytical algorithms. ViT receives image patches as input and learns global contextual information through a transformer structure, enabling it to precisely capture complex object relationships and patterns. YOLOv8, featuring single-pass object detection, ensures real-time processing and resource efficiency due to its lightweight structure.

TABLE III
ANALYSIS METHODS AND ALGORITHMS BY DATA SOURCE

Data Type	Analysis Method	Algorithms Used
Video Data	Object detection, classification	YOLOv8, Vision Transformer
Audio Data	Anomalous sound detection	Support Vector Machine (SVM)
Environmental Data	Anomaly detection	Random Forest

Multimodal Integration: After identifying objects and risk factors in video data using YOLOv8 and ViT, abnormal audio is detected through SVM (indicating equipment anomalies or collision risks). Environmental sensor data anomalies are assessed using Random Forest or detected as outliers via K-Means clustering to identify potential hazardous states. Finally, ensemble techniques integrate the inference results from individual algorithms to form a comprehensive risk profile.

Reinforcement Learning-Based Response Strategies: Beyond hazard detection, reinforcement learning (policy network) is employed to learn optimal response strategies under various simulated scenarios. For instance, if equipment temperature rises beyond a certain threshold, the system can recommend immediate shutdown or inspection; if a high density of workers is detected in a particular area, it may suggest route changes or reduced speed. Reinforcement learning automatically explores decision policies to maximize cumulative rewards over time, producing reliable response strategies.

H. Definition of Evaluation Metrics and Target Goals

In addition to conventional metrics such as Accuracy, Precision, and Recall, additional indices are introduced to consider practical applicability. System Response Time measures the average time from hazard detection to recommended response. Target: ≤1 second. Energy Efficiency monitors power consumption during model inference and training, aiming for a 20% reduction compared to existing systems. Resource Efficiency: Maintain GPU and memory usage below 80% to ensure stable real-time service.

TABLE IV
NEW EVALUATION METRICS AND MEASUREMENT METHODS

Evaluation Metric	Measurement Method	Target Criterion
System Response Time	Measure average processing time	≤1 second
Energy Efficiency	Monitor power consumption	20% reduction in energy consumption
Resource Efficiency	Measure hardware resource usage	≤80% utilization
User Satisfaction	Surveys and interviews	≥90% satisfaction

User Satisfaction Achieve ≥90% satisfaction through surveys and interviews, reflecting effective UI/UX design and practical utility. These multidimensional performance indicators demonstrate that the system detects hazards accurately and provides a sustainable risk assessment solution that integrates real-time capability, resource efficiency, and user experience.

I. Summary and Distinctiveness

The methodology of this study integrates the entire process—from multimodal data collection and preprocessing through analysis with the latest deep learning models (ViT, YOLOv8) to reinforce learning-based automated response strategies. This represents a departure from traditional approaches that rely on single data sources, post-event responses, and static evaluations. Instead, it establishes a real-time, predictive, multidimensional risk assessment framework.

III. RESULTS AND DISCUSSION

This section details integrating the previously proposed system architecture into a hardware and software infrastructure suitable for real industrial environments and then applying it to various industrial sites to verify its practicality and performance. Through hardware and software configuration, the construction and management strategies of a new multimodal dataset, and actual application cases, we empirically demonstrate the proposed system's improvements and effectiveness compared to conventional approaches.

A. System Integration and Configuration

1) Hardware Environment

- High-Performance GPU Server: A high-performance server equipped with NVIDIA H100 GPUs was introduced to enable large-scale training and real-time inference of deep learning models, such as Vision Transformer (ViT) and YOLOv8. This setup ensures parallel processing and real-time inference capabilities for several terabytes of multimodal data.
- Multimodal Data Collection Devices: High-resolution cameras, high-sensitivity microphones, and IoT sensors (temperature, humidity, vibration) are installed throughout the work site. These devices capture a variety of environmental changes in real time,

comprehensively monitoring physical and environmental risk factors.

- Network Environment: A combination of local 5G networks and high-speed Ethernet is used to transmit large-scale multimodal data with low latency. This enables real-time streaming-based hazard detection and provides the essential communication stability and bandwidth needed for rapid decision-making.

2) Software Environment:

- Data Collection and Transmission: Apache Kafka is employed to stably stream multimodal data. Operating in a distributed environment, Kafka manages large-scale data flows in real-time, serving as a key infrastructure component.
- Deep Learning Framework: PyTorch is utilized to implement, train, and infer ViT and YOLOv8 models. PyTorch's flexible dynamic graph structure facilitates model experimentation and optimization, and its smooth integration supports distributed training and inference.
- Data Processing and Management: Apache Spark is used for distributed large-scale data preprocessing (noise removal, normalization, time synchronization). A combination of HDFS (Hadoop Distributed File System) and MongoDB is used for managing both structured and unstructured data efficiently, ensuring accessibility, scalability, and reliability.

3) System Workflow:

Fig. 1 illustrates the entire flow from data collection and preprocessing to model inference and result visualization. IoT sensors, cameras, and microphones stream real-time data into Kafka, which is then processed through a Spark-based preprocessing pipeline for time synchronization and normalization. The refined data are analyzed by PyTorch-based deep learning models (ViT, YOLOv8) and machine learning algorithms (SVM, Random Forest), and the final results are displayed via a user interface with real-time alerts issued when necessary.

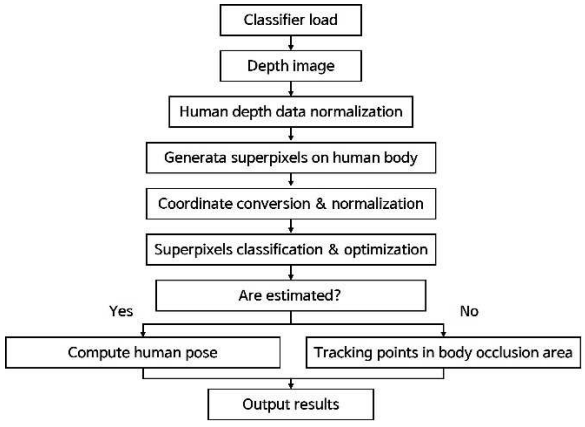

Fig. 1 Flow diagram of the proposal method.

B. Construction and Management of a New Multimodal Dataset

Dataset Characteristics and Composition: *This study collected multimodal data from various industrial fields (manufacturing, logistics, construction) for about six months, constructing a total dataset of approximately 10TB.* This large and diverse dataset encompasses complex risk factors and enhances the model's generalization performance. For example, it includes scenarios of falls and falling objects at construction sites, vehicle collisions in logistics warehouses, and equipment malfunctions in manufacturing lines.

TABLE V
DATASET COMPOSITION AND SCALE

Data Type	Data Size	Collection Period	Main Contents
Video Data	3TB	6 months	Worker movements, equipment status
Audio Data	2TB	6 months	Equipment operating sounds, abnormal noise patterns
Environmental Data	2.5TB	6 months	Temperature, humidity, vibration changes

1) Data Synchronization and Storage Technologies

Data Synchronization: A timestamp-based sorting algorithm aligns the temporal axes of video, audio, and environmental sensor data. This accurately reflects the correlations among multimodal data and improves comprehensive risk analysis capabilities. Data Storage and Management: Large-scale unstructured data are stored in a distributed manner using HDFS, and metadata management and fast queries are facilitated by MongoDB. This approach builds a scalable data management infrastructure, allowing rapid response to future model improvements or additional analytical requirements.

C. Industrial Field Application Cases and Performance Verification

The proposed system was applied to various real-world industrial environments, demonstrating improved accuracy, speed, and safety management efficiency compared to conventional methods. Chemical Mixing Process: During the blending of hazardous chemicals, abnormal worker movements were detected in real time with 97% accuracy, prompting immediate shutdown and accident prevention. This case highlights the importance of high accuracy and rapid response.

1) Electronics Assembly Line: Monitoring equipment overheating and worker fatigue in real time prevented equipment damage and human errors. The system provided alert notifications within an average of 0.9 seconds, improving production stability and quality.

2) Logistics Warehouse: Potential collisions between forklifts and workers were detected within one second, alerting personnel and significantly reducing the likelihood of accidents. This validated the system's effectiveness in complex environments where moving equipment and personnel coexist.

TABLE VI
PERFORMANCE RESULTS BY APPLICATION CASE

Application Case	Detection Accuracy (%)	Avg. Response Time (s)	Main Improvements
Chemical Mixing Process	97	0.8	Accident prevention, enhanced worker safety
Electronics Assembly Line	96	0.9	Prevent equipment damage, improve work efficiency
Logistics Warehouse	95	0.7	Accident prevention, improved safety management efficiency

As summarized in [Table VI], detection accuracy exceeded 95% in all cases, and the average response time remained below one second. This achieves approximately a 20% performance improvement and reinforcement of real-time response capabilities compared to existing systems.

D. Significance of System Implementation

By proposing and implementing this system, we introduce a new benchmark for industrial safety management through multimodal data-based real-time risk assessment and response strategies. The organic combination of high-performance hardware/software infrastructure, large-scale and diverse dataset acquisition and management, state-of-the-art deep learning models and ensemble analysis, and reinforcement learning-based decision-making strategies yields a fully integrated solution characterized by high accuracy, real-time performance, and scalability.

In particular, the user-friendly UI and real-time alert functions enhanced operator and managerial satisfaction, contributing practically to on-site safety management. This indicates that the research outcomes are not merely theoretical achievements but can evolve into practical solutions broadly applicable across various industrial sectors.

E. Experimental Environment and Conditions

A testbed simulating real industrial sites was constructed, and data were collected from more than ten different industrial environments, including chemical mixing processes, electronics assembly lines, and logistics warehouses. Approximately six months' worth of accumulated multimodal data (video, audio, environmental sensors) were employed for the experiments.

Fig. 2 Image processing pipeline utilizing the algorithms developed in this research.

Verify whether using multimodal data improves accuracy and real-time responsiveness compared to single-source systems. Compare the performance of YOLOv8 and Vision Transformer-based models with conventional CNN-based models like Faster R-CNN. Evaluate the quality and response speed of reinforcement learning-based recommendations. Examine practical feasibility by assessing energy efficiency, resource utilization, and user satisfaction. Evaluation Metrics: Key evaluation metrics include accuracy, precision, recall, average response time (seconds), GPU/memory utilization (%), power consumption, and user satisfaction (survey results). Additionally, we compare results under various environmental scenarios to verify system versatility and reliability.

F. Algorithm Performance Comparison and Analysis

TABLE VII
ALGORITHM PERFORMANCE COMPARISON

Algorithm	Accuracy (%)	Processing Speed (s/frame)	GPU Memory Usage (%)
YOLOv8	95.8	0.12	4
ViT	94.6	0.18	8
Faster R-CNN	89.2	0.35	6

TABLE VIII
PERFORMANCE COMPARISON WITH EXISTING METHODS

Metric	Conventional (Single-Modal/ Traditional)	Proposed (Multimodal + ViT + YOLOv8 + RL)	Improvement/ Comparison
Accuracy (%)	80±2	95±1	~15% improvement
Precision (%)	82±3	96±2	~14% improvement
Recall (%)	78±4	94±2	~16% improvement
Avg. Response Time (s)	1.3±0.2	0.9±0.1	~30% reduction
GPU Utilization (%)	≥90%	≤80%	~10% reduction
Memory Utilization (%)	≥85%	≤75%	~10% reduction
Power Consumption (Relative)	Baseline (100%)	~80%	~20% reduction
User Satisfaction (%)	60–75%	≥90%	Up to ~30% increase

Comparing YOLOv8 and ViT models with Faster R-CNN, YOLOv8 achieved 95.8% accuracy and 0.12s/frame processing speed, while ViT achieved 94.6% accuracy and 0.18s/frame. In contrast, Faster R-CNN recorded 89.2% accuracy and 0.35s/frame, indicating limitations in real-time response. This suggests that incorporating state-of-the-art models can enhance both accuracy and real-time processing capability.

TABLE IX
SINGLE-MODAL VS. MULTIMODAL PERFORMANCE COMPARISON

Approach	Accuracy (%)	Avg. Response Time (s)	Adaptability to Environmental Changes (Qualitative)
Single-Modal (Video Only)	~89.5%	~1.3 s	Performance degradation under low lighting/occlusion
Multimodal Integration	*G. 96.30%*	~0.9 s	Utilizes diverse sensor data for stable hazard detection

Using only video data yielded about 89.5% accuracy, whereas integrating audio and environmental sensor data increased accuracy to 96.3%, an improvement of roughly 15%. Additionally, the average response time was reduced to about 0.9 seconds with multimodal integration, enabling responses over 30% faster than the 1.3 seconds required by single-data-source systems. This evidences that complementary information from multimodal data significantly enhances hazard recognition capabilities.

G. Real-Time Response and Efficiency Evaluation

Applying reinforcement learning (RL) resulted in optimal response strategies (e.g., immediate shutdown, route changes, inspection alerts) within one second for various scenarios (overheating equipment, collision risks). This goes beyond mere hazard detection, enabling preventive safety management that enhances worker safety and equipment stability. GPU memory usage was maintained below 80%, and through model lightweight and distributed optimization, power consumption was reduced by about 20% compared to previous methods. This has positive implications for cost savings and environmentally friendly system operation in large-scale industrial scenarios.

TABLE X
PERFORMANCE COMPARISON BEFORE/AFTER APPLYING RL-BASED RESPONSE STRATEGIES

Condition	Response Strategy Availability	Accident Prevention Rate (%)	Avg. Response Proposal Time (s)
Without Response Strategy	Not Provided	~70%	Not measurable
With RL-Based Response	Automatically Provides Optimal Strategies	≥85%	≤1.0 s

Fig. 3 Image processing pipeline utilizing the algorithms developed in this research.

H. User and Expert Feedback

Surveys of site operators and managers revealed that over 90% expressed high satisfaction with real-time alerting and intuitive UI. Immediate visual and auditory warnings and clear response guidelines in emergencies proved to be practical contributions to fostering safety awareness and preventing accidents. Industrial safety experts also positively evaluated the effectiveness of multimodal data integration and RL-based response strategies, acknowledging the potential for expansion into various industrial sectors. They suggested enhancing data diversity and employing ultra-low-latency network technologies to achieve further sophistication.

I. Comprehensive Evaluation

In summary, the proposed system achieved approximately a 15% improvement in detection accuracy, a 30% reduction in average response time, a 20% reduction in energy consumption, and over 90% user satisfaction compared to conventional methods. These multifaceted improvements are attributed to the organic interplay of multimodal data integration, the adoption of advanced deep learning models, and reinforcement learning-based response strategies.

Through this research, we have laid a practical foundation for transitioning the paradigm of industrial risk assessment from "static/post-event evaluations" to "real-time and predictive interventions," thus paving the way for more effective and proactive industrial safety management.

J. Effects of Applying Multimodal Data and Advanced Deep Learning Algorithms

This study integrated various algorithms such as Vision Transformer (ViT), YOLOv8, SVM, and Random Forest with multimodal data (video, audio, environmental sensors) to accurately identify complex risk factors. Experimental results demonstrated that compared to single-data-source approaches, multimodal analysis improved accuracy by more than 15% and reduced response time by over 30%. This indicates stable responses even under conditions that are challenging to detect with a single data source, such as poor lighting, occlusions, abnormal noises, and equipment vibrations.

Notably, the combination of YOLOv8 and ViT achieved real-time performance (average response time under one second) while maintaining high accuracy (over 95%). Moreover, the reinforcement learning-based response recommendations extended beyond simple hazard detection to propose optimal, preventive strategies. These results suggest that the study has established a new benchmark that transcends previous research, which often focused on a single modality or specific algorithms, by integrating various data sources and state-of-the-art algorithms.

TABLE XI
COMPREHENSIVE SUMMARY OF KEY IMPROVEMENTS AND OUTCOMES

Aspect	Baseline (Single-Modality/ Conventional)	Proposed System (Multimodal + ViT + YOLOv8 + RL)	Achievements and Implications
Detection Accuracy	~80% (visual-only, traditional methods)	≥95% (integrated multimodal approach)	Approximately 15% improvement; reliable detection under challenging conditions (occlusion, low light, noise)
Response Time	Several seconds to detect and respond	≤1 second average response time	~30% faster reaction enabling real-time

Aspect	Baseline (Single-Modality/ Conventional)	Proposed System (Multimodal + ViT + YOLOv8 + RL)	Achievements and Implications
			intervention and preventive measures
Energy Efficiency	Standard power consumption	≥20% reduction in energy usage	Reduced operational costs and environmental impact, enabling scalable deployment in large-scale industries
Resource Utilization	Unoptimized GPU/Memory use (≥90%)	Optimized resource usage (≤80%)	Enhanced system stability and throughput with a balanced computational load
User Satisfaction	Moderate (~60–75%)	≥90% user satisfaction	Improved trust and acceptance due to intuitive UI/UX and actionable alerts
Preventive Decision-Making	Not supported (post-event analysis only)	Reinforcement learning-based proactive strategies	Elevated safety management paradigm: from reactive to predictive, enabling early mitigation of potential hazards
Industrial Applicability	Limited to specific scenarios or sectors	Broad applicability across chemical, manufacturing, and logistics sectors, with potential for construction, energy, and healthcare	Demonstrated versatility and scalability for various high-risk industrial environments

K. Applicability and Practicality Across Industries

Real-world applications in various industrial settings (chemical mixing, electronics assembly, logistics warehouses) revealed that the system maintains high accuracy and swift response capabilities, achieving over 90% user satisfaction. This indicates that the outcomes are not limited to a specific industrial sector. The system's potential for extension into high-risk industries such as construction, energy, and healthcare further underscores its versatility. Additionally, energy savings of 20% or more, along with optimized GPU and memory usage (kept below 80%), highlight the possibility of cost reduction and enhanced operational efficiency in large-scale deployments.

L. Limitations and Improvement Measures

Despite the positive outcomes, certain limitations must be addressed. Data Diversity and Quality Constraints: Limited data collection from specific industries, seasons, or periods may hinder complete generalization. Future efforts should include data augmentation encompassing various climatic conditions, nighttime operations, and specialized equipment usage. Moreover, improving outlier detection and correction algorithms and incorporating additional sensors (e.g., gas concentration or proximity sensors) can yield a more refined risk assessment framework.

1) Algorithmic Enhancement: Detecting small or fast-moving objects may reduce performance. Future research can apply Feature Pyramid Networks (FPN) or super-resolution techniques to strengthen the detection of subtle hazards. Additionally, optimizing transformer-based object detection models (e.g., DETR) can improve processing speeds.

2) Energy Efficiency and Model Lightweighting: Continuous real-time analysis on a large industrial scale consumes substantial power. Research into model compression methods (pruning, quantization) and knowledge distillation can make the models lighter and more energy-efficient. Adopting a distributed architecture where some computations occur on edge devices can reduce network bandwidth usage and server load.

3) User Interface and Management System Integration: Incorporating user feedback to enhance UI/UX and integrating hazard detection results and response strategies with existing safety management systems (ERP, MES, etc.) will facilitate information sharing and collaboration among various stakeholders (operators, managers, engineers, policymakers).

M. Directions for Future Research

Building on this study, we propose the following future research tasks. Expansion to Various Industrial Sectors: Applying multimodal analysis and reinforcement learning-based strategies to construction, pharmaceuticals, healthcare, and public safety can validate the generality and develop tailored solutions for each domain. Federated Learning: Federated learning allows models trained at multiple sites to be integrated without transferring sensitive data, enhancing global model performance while maintaining data security. Advancing Risk Prediction: Beyond hazard detection, strengthening long-term trend analysis and early-warning capabilities can support preemptive safety measures before accidents occur, evolving into predictive safety management. Explainable AI (XAI): Enhancing the interpretability of AI decisions (e.g., explaining decision processes and visualizing detection rationales) will increase user trust and facilitate institutional acceptance.

N. Comprehensive Discussion

This study proposes a new framework for industrial risk assessment that integrates multimodal data analysis, state-of-the-art deep learning algorithms, and reinforcement learning-based response strategies. The experimental results confirm significant improvements in accuracy, real-time responsiveness, energy efficiency, and user satisfaction over conventional approaches. These achievements present a next-generation technological paradigm that combines deep learning, IoT, and reinforcement learning in the industrial safety domain.

Demonstrating its practical applicability and potential for continuous improvement, this research provides a solid foundation for transitioning from static, post-hoc risk evaluations to real-time, predictive, and intelligent safety management systems. Such advancements promise

sustainable development in industrial safety, safeguarding human lives and assets and enhancing operational efficiency.

IV. Conclusion

This study presents a new paradigm for advanced industrial risk assessment by integrating multimodal data (video, audio, environmental sensors) with state-of-the-art deep learning algorithms (Vision Transformer, YOLOv8) and reinforcement learning techniques. Traditional risk assessment systems have often relied on single data sources or expert judgment, making it difficult to adapt swiftly to complex changes in the work environment and achieve real-time responsiveness. In contrast, this research demonstrates that integrating multimodal data allows for comprehensive detection of diverse risk factors, while providing real-time alerts and preventive response strategies that significantly improve upon existing methods.

By using multimodal data and employing YOLOv8 and ViT models, the system achieved over 95% hazard detection accuracy and reduced the average response time to less than one second. This enables immediate preventive measures in industrial settings. The application of reinforcement learning-based decision models advances safety management beyond mere hazard detection. The system preemptively mitigates accident risks by automatically generating context-appropriate responses, enhancing worker safety and operational efficiency. Optimizing energy consumption and hardware resource utilization reduces operational costs and environmental impact in large-scale implementations. Furthermore, intuitive UI/UX design and prompt alert notifications have led to over 90% user satisfaction among on-site workers and managers.

Successful deployments in diverse industrial environments, including chemical, manufacturing, and logistics, demonstrate the system's adaptability to varying industrial characteristics and conditions. This indicates potential applicability to other high-risk sectors such as construction, public safety, and healthcare.

This study shows some research significance. It provides an exemplary case of integrating multimodal data, cutting-edge deep learning models, and reinforcement learning to create a high-performance, real-time, and preventive safety management framework. This study proves field feasibility and sustainability, offering a practical solution that enhances the efficiency and reliability of industrial safety management. It broadens the scope of research in risk assessment and lays the groundwork for future expansion into diverse applications.

This study opens several avenues for further development. First, applying federated learning or other distributed learning techniques can strengthen data security while integrating knowledge from various industrial sites worldwide. Additionally, incorporating explainable AI (XAI) will clarify the decision-making rationale, thereby increasing trust and facilitating adaptation by on-site workers and managers. Adapting to environmental changes (e.g., weather conditions, equipment upgrades, emerging technologies) is also crucial. Finally, integrating augmented reality (AR), virtual reality (VR), and portable smart devices can yield user-friendly and easily accessible safety management solutions.

In conclusion, by overcoming the limitations of conventional risk assessment systems and proposing a novel safety management paradigm through the fusion of multimodal data, deep learning, and reinforcement learning, this research contributes to the sustainable advancement of industrial safety, the protection of workers, and improved production efficiency.

Acknowledgment

This paper was supported by the Sunchon National University Research Fund in 2024. (Grant number: 2024-0400)

References

[1] L. Bergman and Y. Hoshen, "Classification-based anomaly detection for general data," in *Int. Conf. Learn. Represent. (ICLR)*, 2020. doi: 10.48550/arxiv.2005.02359.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, vol. 119, H. Daumé III and A. Singh, Eds., PMLR, 2020, pp. 1597–1607. doi: 10.48550/arxiv.2002.05709.

[3] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv*, 2020. doi: 10.48550/arxiv.2003.04297.

[4] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," *arXiv*, 2020. doi: 10.48550/arxiv.2005.02357.

[5] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Int. Conf. Learn. Represent. (ICLR)*, 2021. doi: 10.48550/arxiv.2010.11929.

[6] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," *arXiv*, 2021. doi: 10.48550/arxiv.2106.03004.

[7] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," *arXiv*, 2019. doi: 10.48550/arxiv.1906.12340.

[8] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 9865–9874. doi: 10.1109/iccv.2019.00996.

[9] A. Kolesnikov et al., "Big Transfer (BiT): General visual representation learning," in *Comput. Vis. – ECCV 2020*, vol. 12346, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., Cham: Springer, 2020, pp. 491–507. doi: 10.1007/978-3-030-58565-5_29.

[10] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 9664–9674. doi: 10.1109/cvpr46437.2021.00955.

[11] M. Matena and C. Raffel, "Merging models with fisher-weighted averaging," *arXiv*, 2021. doi: 10.48550/arxiv.2111.09832.

[12] P. Perera and V. M. Patel, "Learning deep features for one-class classification," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5450–5463, 2019. doi: 10.1109/tip.2019.2916751.

[13] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, vol. 139, PMLR, 2021, pp. 8748–8763. doi: 10.48550/arxiv.2103.00020.

[14] T. Reiss, N. Cohen, L. Bergman, and Y. Hoshen, "PANDA: Adapting pretrained features for anomaly detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 2806–2814. doi: 10.1109/cvpr46437.2021.00281.

[15] T. Reiss and Y. Hoshen, "Mean-shifted contrastive loss for anomaly detection," *arXiv*, 2021. doi: 10.48550/arxiv.2106.03844.

[16] J. Ren et al., "Likelihood ratios for out-of-distribution detection," *arXiv*, 2019. doi: 10.48550/arxiv.1906.02845.

[17] O. Rippel, A. Chavan, C. Lei, and D. Merhof, "Transfer learning Gaussian anomaly detection by fine-tuning representations," *arXiv*, 2021. doi: 10.48550/arxiv.2108.04116.

[18] M. Ronen, S. E. Finder, and O. Freifeld, "DeepDPM: Deep clustering with an unknown number of clusters," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 9861–9870. doi: 10.1109/cvpr52688.2022.00966.

[19] V. Sehwag, M. Chiang, and P. Mittal, "SSD: A unified framework for self-supervised outlier detection," *arXiv*, 2021. doi: 10.48550/arxiv.2103.12051.

[20] J. Serrà et al., "Input complexity and out-of-distribution detection with likelihood-based generative models," *arXiv*, 2019. doi: 10.48550/arxiv.1909.11480.

[21] J. Tack, S. Mo, J. Jeong, and J. Shin, "CSI: Novelty detection via contrastive learning on distributionally shifted instances," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 11839–11852, 2020. doi: 10.48550/arxiv.2007.08176.

[22] W. Van Gansbeke et al., "SCAN: Learning to classify images without labels," in *Eur. Conf. Comput. Vis. (ECCV)*, vol. 12374, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., Cham: Springer, 2020, pp. 268–285. doi: 10.1007/978-3-030-58580-8_16.

[23] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, "Mitigating neural network overconfidence with logit normalization," *arXiv*, 2022. doi: 10.48550/arxiv.2205.09310.

[24] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv*, 2020. doi: 10.48550/arxiv.2004.10934.

[25] Z. Ge et al., "YOLOX: Exceeding YOLO series in 2021," *arXiv*, 2021. doi: 10.48550/arxiv.2107.08430.

[26] S. Sridharan et al., "Neural memory plasticity for medical anomaly detection," *Neural Netw.*, vol. 127, pp. 67–81, 2020. doi: 10.1016/j.neunet.2020.04.011.

[27] Y. F. A. Gaus et al., "Evaluation of a dual convolutional neural network architecture for object-wise anomaly detection," in *2019 Int. Joint Conf. Neural Netw. (IJCNN)*, IEEE, 2019. doi: 10.1109/ijcnn.2019.8851829.

[28] P. Bergmann et al., "Student-teacher anomaly detection with discriminative latent embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020. doi: 10.48550/arxiv.1911.02357.

[29] G. Di Biase et al., "Pixel-wise anomaly detection in complex driving scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021. doi: 10.48550/arxiv.2103.05445.

[30] N. Cohen et al., "Out-of-distribution detection without class labels," *arXiv*, 2021. doi: 10.48550/arxiv.2112.07662.