Intelligent Systems and Robotics

E-ISSN: xxxx-xxxx
P-ISSN: xxxx-xxxx

*Research Article*

# Real Time Computer Vision System Based on Convolutional Neural Networks for Precision Object Detection and Tracking in Collaborative Industrial Robot Applications

**Anggit Wirasto [1], Khoirun Nisa [2], Titi Christiana [3]**

1 Universitas Harapan Bangsa anggitwirasto@uhb.ac.id
2 Universitas Harapan Bangsa khoirunnisa@uhb.ac.id
3 Universitas Sains dan Teknologi kumputer titi@stekom.ac.id
* Corresponding Author : Anggit Wirasto

**Abstract:** The increasing adoption of collaborative robots in modern manufacturing environments requires reliable perception systems that can ensure both safety and operational efficiency during human–robot collaboration. This study proposes a CNN-based real-time computer vision system for object and human detection in shared robotic workspaces. The research focuses on developing and evaluating a single-stage deep learning detection model optimized for real-time performance while maintaining high detection accuracy. The proposed methodology includes dataset preparation, model training using transfer learning, real-time system implementation, and comprehensive performance evaluation. Experimental results demonstrate that the developed system achieves high detection accuracy, as reflected by strong precision, recall, and mean Average Precision (mAP) values, while maintaining low inference latency suitable for real-time operation. The system consistently operates above real-time frame-rate thresholds, ensuring timely perception updates required for safety-related decision-making in collaborative robotic environments. Graphical and quantitative analyses further confirm the stability of inference performance under dynamic interaction scenarios involving human movement and multiple objects. Compared with existing approaches, the proposed system provides a balanced trade-off between accuracy and computational efficiency, making it practical for deployment in safety-aware human–robot collaboration scenarios. Overall, the findings indicate that CNN-based real-time object detection systems can effectively support perception and situational awareness in collaborative robotics, contributing to safer and more efficient industrial automation.

**Keywords:** Collaborative robotics; Computer vision; Convolutional neural networks; Object detection; Real-time systems.

**Received: November 20, 2025**
**Revised: Desember 30, 2025**
**Accepted: January 14, 2026**
**Published: January 18, 2026**
**Curr. Ver.: January 20, 2026**

**Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)**

## 1. Introduction

The rapid evolution of manufacturing systems in the context of Industry 4.0 and the emerging paradigm of Industry 5.0 has intensified the adoption of collaborative industrial robots, commonly referred to as cobots. Unlike traditional industrial robots that operate in isolated and fenced environments, cobots are specifically designed to share workspaces with human operators, enabling direct physical and cognitive collaboration. This paradigm shift aims to enhance flexibility, productivity, and safety while preserving the central role of human workers in manufacturing processes [1], [2]. One of the primary advantages of collaborative robots lies in their flexibility and ease of deployment. Cobots are typically characterized by intuitive programming, reconfigurability, and adaptability to diverse manufacturing tasks such as assembly, pick-and-place operations, packaging, and quality inspection. These characteristics make them particularly suitable for small-batch and high-mix production

environments, where conventional automation solutions are often economically or technically infeasible [1], [3].

Safety is a fundamental pillar of human robot collaboration. Cobots integrate multiple safety mechanisms, including power and force limiting, speed and separation monitoring, collision detection, and ergonomic interaction design, to mitigate risks associated with close human–robot proximity. Extensive research has demonstrated that these mechanisms significantly reduce the likelihood of severe injuries while maintaining operational efficiency [4], [5]. However, ensuring safe collaboration remains a complex challenge, particularly in dynamic and unstructured manufacturing environments where human behavior is unpredictable [1], [6]. In addition to safety considerations, collaborative robots have been shown to improve overall manufacturing efficiency and productivity. By working alongside human operators, cobots can reduce cycle times, improve process consistency, and enhance product quality while allowing workers to focus on higher-value cognitive and decision-making tasks. Empirical studies indicate that such synergies contribute to both operational performance and worker satisfaction when properly implemented [3], [7].

Despite their benefits, the implementation of collaborative robots presents several challenges. High initial investment costs, integration with existing infrastructure, and cybersecurity concerns remain critical barriers, particularly for small and medium-sized enterprises [6], [8]. Furthermore, the impact of cobots on workforce skills is a growing concern. While some studies report reskilling and upskilling opportunities, others highlight risks of task simplification and deskilling depending on the level of autonomy and interaction design [9]. Recent research trends emphasize the integration of advanced technologies such as artificial intelligence, machine learning, multimodal sensing, and augmented reality to enhance cobot perception, adaptability, and decision-making capabilities. These technologies enable more natural, context-aware, and intuitive human–robot interactions, supporting safer and more efficient collaboration [7], [10], [11]. Moreover, human-centered and cognitively ergonomic design approaches are increasingly adopted to align collaborative robotic systems with human capabilities, limitations, and well-being [12].

In line with the principles of Industry 5.0, future developments in collaborative robotics are expected to prioritize human-centricity, sustainability, and resilience. Cobots are envisioned not merely as productivity-enhancing tools but as partners that support inclusive, adaptive, and environmentally responsible manufacturing systems [2], [7]. Consequently, a comprehensive understanding of the technological, safety, and socio-economic dimensions of collaborative robots is essential to fully realize their potential in modern manufacturing environments.

## 2. Literature Review

### Overview of Vision-Based Human Robot Collaboration (HRC)

Computer vision has become a core enabling technology for human–robot collaboration (HRC) in modern manufacturing because it provides real-time situational awareness of human presence, posture, motion, and workspace context. Vision-based perception supports barrierless collaboration by allowing robots to perceive dynamic environments without relying solely on physical fences or simple proximity sensors [13]. Recent reviews emphasize that computer vision often combined with AI has shifted HRC from static, pre-programmed coordination toward adaptive collaboration, where robots can react to human behavior and task state with higher autonomy [14]. In smart manufacturing, this trend aligns with broader digitalization efforts (e.g., data-driven decision making and digital-twin ecosystems) that demand robust perception pipelines for safe and efficient automation [15].

## Computer Vision for Safety: Collision Avoidance and Hazard Prevention

### Human detection, 3D tracking, and collision avoidance

A primary safety contribution of computer vision in cobot environments is collision risk reduction through detection and tracking of human motion. Vision-based safety systems can estimate worker location and movement in real time, enabling robots to adjust trajectories, reduce speed, or stop when humans enter defined safety zones [16]. A practical approach that has gained attention is skeleton-based tracking, where human joint landmarks are tracked continuously to infer body pose and proximity, supporting collision avoidance decisions that are sensitive to human movement patterns [17]. These approaches are important for open, shared workspaces because safety must be maintained despite variability in human behavior and task flow [13], [18].

### Speed and Separation Monitoring (SSM) using vision

Vision-based implementations of speed and separation monitoring (SSM) operationalize safety by linking robot speed control to measured human–robot distance. As the closest human approaches, robot speed can be reduced to ensure stopping time remains within safe limits, thereby supporting barrierless collaboration [16]. Prior work in vision-based safety notes that such systems require reliable perception (e.g., accurate human localization and low-latency processing), because safety margins can be compromised by sensor noise, occlusions, or delayed detection [13]. More recent fusion approaches address these limitations by combining multiple vision streams or integrating complementary sensing and AI reasoning to improve robustness under real factory-floor conditions [18].

### Vision enabled real time hazard analysis in complex tasks

Beyond immediate collision avoidance, computer vision can support higher-level safety reasoning via job hazard analysis (JHA). In disassembly settings, where tools, parts, and worker actions can change rapidly, vision-enabled real-time hazard analysis can identify unsafe states (e.g., risky proximity to moving robot elements or hazardous part-handling moments) and provide control interventions or warnings [19]. Such approaches extend safety from distance-based rules toward context-aware risk assessment, which is increasingly necessary for complex collaborative tasks involving variable objects and workspace layouts [14].

## Efficiency Gains from Vision Based Perception in Collaborative Robotics

### Reducing unnecessary stops and improving workflow continuity

Efficiency improvements arise when vision systems can distinguish humans from non-human objects and interpret motion intent, reducing overly conservative robot behaviors. If the robot can accurately classify approaching objects and estimate whether they represent a true collision risk, it can avoid unnecessary slowdowns or stoppages that degrade throughput [18]. In smart manufacturing, where cycle-time efficiency and flexibility are crucial, vision-based perception becomes an operational tool for maintaining productivity while preserving safety constraints [15].

### Natural interaction: gesture, tool handover, and intention estimation

A second efficiency pathway is enabling more natural interaction patterns, including gesture-based control and tool handover. Computer vision can interpret human gestures or hand poses to trigger robot actions, allowing faster and more intuitive coordination than explicit programming or manual interface use [20]. Closely related is intention estimation, where vision-based posture understanding helps robots anticipate human actions, improving fluency and reducing idle time caused by uncertain coordination. Although demonstrated strongly in construction contexts, posture-based intention estimation is conceptually transferable to manufacturing HRC, particularly for tasks involving close physical collaboration [21]. Together, gesture and intention awareness can improve task allocation, reduce communication overhead, and enhance responsiveness in dynamic workflows [20].

### Adaptability and responsiveness through AI-enhanced vision

AI-enhanced computer vision strengthens adaptability by enabling learning-based detection and decision policies that generalize across lighting changes, occlusions, and diverse worker behaviors. Reviews highlight that combining AI and vision is central to next-generation collaborative robotics, improving perception accuracy and expanding the range of feasible collaborative scenarios [14]. In parallel, smart manufacturing roadmaps position computer vision as a key component for responsive robotic systems that can reconfigure quickly and operate robustly within data-driven production environments [15].

## Key Challenges and Emerging Solutions

### Reliability under occlusion, latency constraints, and sensor fusion

Manufacturing environments can be visually challenging due to clutter, reflective surfaces, variable illumination, and frequent occlusion by workers or objects. Vision-based safety systems must therefore be designed with robustness and latency constraints in mind, as delays or misdetections directly affect safety margins [13]. Sensor fusion and vision fusion strategies have been proposed to mitigate these risks by combining complementary information sources and improving confidence in human localization and hazard detection [18]. Research also increasingly frames these solutions as part of broader AI–vision integration to improve resilience and reliability [14].

### Data privacy and security implications of factory-floor vision

The increased deployment of cameras introduces governance concerns, including privacy, worker acceptance, and data protection. While the provided literature emphasizes technical fusion and safety management, the need for responsible data handling grows as vision becomes more pervasive and integrated into production analytics [15], [18]. This motivates approaches that minimize personally identifiable data, enforce secure processing pipelines, and define transparent policies for workforce stakeholders.

### Standardization and scalable deployment

Another persistent challenge is the limited standardization of vision-based safety and interaction solutions, which slows adoption and interoperability across heterogeneous equipment and factory settings. Early reviews already highlighted the need for systematic evaluation methods and safety-aligned design practices for vision-based HRC (Halme et al., 2018). Current work continues to stress that scalable deployment requires not only better algorithms but also repeatable integration procedures, reliable validation protocols, and cross-system compatibility [14], [15].

## Synthesis and Research Direction

Across the literature, a consistent finding is that computer vision significantly strengthens both safety and efficiency in collaborative robotics by enabling real-time awareness of human state, workspace context, and task dynamics. The safety dimension is advanced through collision avoidance, SSM implementations, and context-aware hazard analysis [16], [17], [19] . Efficiency is improved via reduced unnecessary downtime, more natural gesture- and handover-based interaction, and AI-enhanced adaptability [15], [20]. However, real-world deployment remains constrained by perception reliability, latency, privacy concerns, and the need for standardization [13], [18]. These gaps motivate continued research on robust, low-latency perception; fusion-based safety assurance; and scalable frameworks that balance productivity, worker well-being, and responsible data practices in barrierless collaborative workspaces [14].

### Foundations of Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a class of deep neural architectures designed to process grid-structured data most notably images by learning hierarchical feature representations directly from pixel-level input. The core principle of CNNs is locality: convolution operations apply trainable filters over local receptive fields, enabling the model to detect low-level patterns (e.g., edges and textures) and progressively compose them into higher-level semantic features [22]. This layered feature learning makes CNNs particularly effective for image recognition and computer vision tasks compared with earlier feature-engineering approaches, because the model can optimize feature extraction and classification jointly [23].

A canonical CNN pipeline consists of stacked convolutional layers, typically interleaved with nonlinear activation functions and downsampling mechanisms (e.g., pooling) to improve invariance to small translations and reduce computational load. Convolutional layers transform the input image into feature maps that capture spatial patterns relevant to object identity and appearance [22]. At later stages, fully connected layers (or equivalent classification heads) integrate extracted features to produce class probabilities, although modern detectors often replace heavy fully connected blocks with more parameter-efficient heads [23]. Across applications, CNNs have proven robust for representation learning, serving as backbones for classification, detection, segmentation, and medical image analysis [24].

### Evolution of Deep Learning Object Detection Paradigms

Object detection extends recognition by requiring models to localize objects (e.g., bounding boxes) while also classifying them. Historically, CNN-based object detection advanced through two dominant paradigms: (1) region-based, two-stage detectors and (2) single-stage detectors optimized for speed.

### Two-stage detectors: R-CNN family and region proposals

Two-stage detectors first generate candidate regions likely to contain objects and then classify/refine these proposals. This approach became influential because it improved localization quality and detection accuracy relative to earlier sliding-window methods. Surveys of object detection research commonly categorize R-CNN-style methods as foundational, because they demonstrated how CNN features could support both localization and classification within a unified learning framework [25]. In practice, two-stage detectors are often associated with stronger accuracy under complex backgrounds but higher computational cost, making them less suitable for strict real-time requirements when compared to single-stage approaches [25].

### Single-stage detectors: YOLO and real-time detection emphasis

Single-stage detectors reframed detection as a direct regression/classification problem over dense predictions, enabling fast inference. YOLO-family models are frequently highlighted for real-time performance because they predict bounding boxes and class probabilities in a single forward pass [26]. Comparative work focusing on modern YOLO variants shows that later versions can improve accuracy while preserving speed, making them attractive for deployment in time-sensitive environments [27]. Real-time object detection studies further show that the practicality of deployment depends not only on mean accuracy but also on latency, throughput, and robustness under operational conditions such as varying illumination and scene dynamics [28].

### CNNs and Transformers in Contemporary Detection

Recent research has expanded beyond purely CNN-based backbones by introducing transformer-based architectures for vision tasks. Surveys examining CNN- and transformer-based detectors report that transformer models can achieve competitive and sometimes superior results by modeling long-range dependencies more effectively than standard convolutions [25]. This shift is relevant for detection scenarios where global context helps

disambiguate objects under occlusion or clutter. However, CNNs remain widely used due to their computational efficiency and mature deployment ecosystem, especially for real-time applications where inference cost and hardware constraints dominate design choices (Cohen et al. not provided here; use only given sources) [9].

### Real-Time Object Detection: Performance Drivers and Evaluation

Real-time detection systems are evaluated using a combination of accuracy metrics (e.g., precision/recall or benchmark scores) and speed metrics (e.g., frames per second and end-to-end latency). Studies on real-time detection emphasize that single-stage methods generally provide superior speed while remaining sufficiently accurate for many operational contexts [26]. For example, comparative analyses of YOLOv3 versus YOLOv7 in OpenCV-oriented pipelines illustrate how detector selection can be optimized based on the trade-off between accuracy improvements and computational requirements [27]. In addition, research on small-target scenarios highlights that detection performance can degrade substantially when objects occupy only a small number of pixels, requiring careful model choice and evaluation tailored to the target domain [29].

Hardware efficiency is another critical driver. When object detection is deployed on embedded systems or constrained industrial platforms, optimized implementations using specialized accelerators (e.g., VLSI-oriented designs) and efficient deep learning pipelines are frequently discussed as necessary conditions for practical real-time performance [30]. This aligns with broader smart manufacturing needs, where detection pipelines must sustain real-time response while operating under limited power, compute budgets, and reliability requirements [30].

### Application Domains Emphasizing Real-Time Detection

Real-time object detection is widely adopted in surveillance and safety monitoring, where the goal is to detect threats or anomalous objects quickly enough to trigger timely interventions. Recent work demonstrates how deep learning detectors can support surveillance pipelines for continuous monitoring, reinforcing the importance of low-latency inference and stable detection performance under noisy visual conditions [28] . Medical imaging represents another domain where CNN-based detection and classification are critical; however, this domain often prioritizes diagnostic accuracy and robustness over real-time constraints. Nevertheless, CNN advances in medical imaging such as improved feature learning and transfer learning also inform broader detection research by introducing techniques for handling limited data, class imbalance, and domain shifts [24], [31].

### Synthesis and Implications for Real-Time Vision Systems

Overall, the literature indicates that CNNs remain central to object detection due to their efficient feature extraction and strong performance across diverse tasks. The evolution from region-proposal methods to single-stage detectors reflects an increasing emphasis on real-time deployment, where speed and resource consumption are critical design constraints [25], [26]. Meanwhile, transformer-based approaches represent a complementary direction that can enhance contextual reasoning, though adoption in real-time systems must account for computational overhead and deployment feasibility [25]. Finally, application-driven evaluation especially under small-target conditions and constrained hardware reinforces the need to align model selection, optimization, and benchmarking with real operational requirements [29], [30].

### Artificial Intelligence and Computer Vision in Industrial Automation

The development of smart industrial systems increasingly relies on artificial intelligence (AI) to enable machines to perceive and understand their environment. One of the most important technologies supporting this capability is computer vision, which allows machines to interpret visual information captured from cameras or sensors. In modern manufacturing environments, computer vision is widely applied in robotic systems to identify objects, monitor processes, and support automated decision-making. The integration of AI within

industrial systems contributes to improved operational efficiency, accuracy, and adaptability in dynamic environments. AI-based frameworks have been shown to strengthen system intelligence by enabling machines to process large volumes of data and generate adaptive responses in real time [32], [33].

In collaborative industrial environments, robots are increasingly required to work alongside human operators. This collaboration requires robots to possess high levels of perception and environmental awareness to ensure safety and precision in task execution. Real-time computer vision systems therefore play a crucial role in enabling robots to detect and interpret objects, movements, and spatial relationships within their operational workspace. The development of intelligent systems supported by AI technologies provides a strong foundation for implementing advanced perception mechanisms in collaborative robotics, thereby enhancing productivity and operational safety.

**Convolutional Neural Networks for Visual Feature Extraction**

Convolutional Neural Networks (CNNs) are widely recognized as one of the most effective deep learning architectures for image processing and visual recognition tasks. CNNs are capable of extracting hierarchical features from images, starting from simple edge detection to more complex representations such as shapes, textures, and object structures. This hierarchical feature extraction enables CNN-based models to achieve high accuracy in object classification, detection, and recognition tasks within complex environments.

Several studies have demonstrated the effectiveness of hybrid CNN models in processing large-scale data and identifying patterns in real-time systems. Research on hybrid CNN architectures integrated with other neural network models has shown promising results in pattern recognition and early detection systems operating in dynamic environments [34], [35]. Furthermore, systematic reviews on emerging technologies highlight the role of intelligent learning models in improving the performance of data-driven systems by enabling efficient feature extraction and pattern analysis [35].

In the context of collaborative industrial robots, CNN-based computer vision systems are particularly useful for enabling robots to recognize objects with high precision. The ability of CNNs to process image data quickly and accurately supports the development of real-time object detection and tracking mechanisms, which are essential for robotic systems operating in continuously changing industrial environments.

**Real Time Object Detection and Tracking Systems**

Object detection and tracking are fundamental components of real-time computer vision systems. Object detection refers to the ability of a system to identify and classify objects within an image or video stream, while object tracking focuses on continuously monitoring the movement of these objects over time. In collaborative industrial robotics, both capabilities are necessary to ensure that robots can interact safely and accurately with objects and human operators.

The implementation of real-time intelligent systems often requires adaptive computational architectures capable of handling high-speed data processing. Distributed and federated learning approaches have been proposed to support scalable real-time analysis in complex computing environments. Studies on federated and ensemble learning approaches demonstrate how distributed intelligence can enhance system responsiveness and maintain performance under dynamic conditions [36]. Similarly, hybrid deep learning architectures combining multiple neural network models have been developed to improve detection accuracy and adaptability in real-time environments [35].

These developments highlight the importance of integrating advanced learning models with real-time processing frameworks to achieve reliable object detection and tracking. In collaborative robot applications, such systems allow robots to continuously monitor object positions and movements, enabling more accurate and responsive interactions with their surroundings.

### Integration of Sensors, IoT, and Intelligent Systems

The effectiveness of real-time computer vision systems also depends on the integration of sensors, data acquisition devices, and communication networks. Modern industrial environments often adopt Internet of Things (IoT) technologies to enable interconnected devices to collect, transmit, and process data in real time. The integration of sensors and intelligent systems provides the infrastructure necessary for developing adaptive and responsive automation solutions.

Several studies have demonstrated the application of IoT-based systems in monitoring, automation, and environmental sensing. For example, IoT technologies have been applied in monitoring water quality and environmental conditions in urban areas, illustrating how sensor networks can continuously collect data and support intelligent analysis [37]. Similarly, IoT-based security systems integrating RFID and sensor technologies demonstrate how hardware components can collaborate with software intelligence to create responsive monitoring systems [38]. Automation systems built on microcontroller platforms further illustrate the role of sensor integration in enabling real-time operational monitoring and control [39].

In industrial robotics, camera sensors serve as the primary input devices for computer vision systems. When combined with IoT connectivity and intelligent processing models, these sensors enable robots to continuously capture visual data and analyze it for object detection and tracking. This integration forms the technological foundation of modern smart manufacturing systems.

### Security and Reliability in Intelligent Industrial Systems

As industrial systems become increasingly connected and data-driven, ensuring system security and operational reliability becomes a critical concern. Real-time vision systems operating in industrial environments must be protected against potential cyber threats and operational disruptions that may affect system stability or safety.

Research on advanced security architectures emphasizes the importance of implementing adaptive security frameworks capable of protecting digital infrastructures from intelligent attacks. Approaches such as blockchain-based security mechanisms, zero-trust architectures, and secure cloud-edge frameworks have been proposed to improve system resilience and ensure continuous service availability [33], [37]. In addition, systematic reviews on cybersecurity technologies highlight the importance of proactive defense strategies to mitigate emerging digital threats in modern computing environments [35].

Although these studies primarily focus on cybersecurity systems, the underlying principles of security and reliability are equally relevant for intelligent robotics systems. Collaborative robots equipped with computer vision capabilities must operate within secure and reliable infrastructures to ensure safe interactions with human operators and industrial equipment.

### Human Technology Interaction in Smart Systems

The successful implementation of intelligent technologies also depends on how effectively humans interact with these systems. In collaborative robotic environments, robots are designed to assist human workers rather than replace them, emphasizing the importance of intuitive and safe human machine interaction.

Research on technology-based learning systems and digital engagement highlights how intelligent technologies can enhance user interaction and system usability. For instance, studies on gamification and digital engagement demonstrate that technology adoption is strongly influenced by user experience and system accessibility [40]. Similarly, innovations integrating technology with learning and practical activities illustrate how digital systems can support interactive and collaborative environments [41], [42].

In industrial robotics, these principles translate into the design of collaborative systems that enable robots and humans to work together efficiently. Computer vision systems capable of accurately detecting and tracking objects contribute to safer interactions between robots and human operators, supporting the broader goals of smart manufacturing and Industry 4.0.

## 3. Research Method

### Research Design

This study adopts an experimental and system development research design focusing on the development, implementation, and evaluation of a real-time computer vision system based on Convolutional Neural Networks (CNNs) for object detection in collaborative robotic environments. The research aims to evaluate the capability of deep learning–based object detection models to support safe and efficient human–robot collaboration (HRC) by achieving high detection accuracy with low latency.

The methodology is structured into sequential phases, including dataset preparation, model selection and training, system implementation, and performance evaluation under real-time constraints.

### System Architecture

The proposed system architecture consists of four main components. The Vision Acquisition Module employs an RGB camera to capture real-time image streams from the collaborative workspace. The camera is strategically positioned to cover the shared human–robot interaction area, ensuring an adequate field of view for detecting humans and relevant objects. The CNN-Based Object Detection Module utilizes a deep learning–based object detection model to perform real-time detection. Single-stage detectors, particularly YOLO-based architectures, are selected due to their favorable balance between detection accuracy and inference speed, making them well suited for real-time robotic applications. The detection results are then processed by the Decision and Safety Logic Module to estimate the spatial relationship between detected humans and the robot workspace. Based on predefined safety thresholds, this module generates appropriate safety-related decisions, such as warning signals, robot speed reduction, or emergency stop commands. Finally, the Evaluation and Monitoring Module records key performance metrics, including detection accuracy, inference time, frame rate, and system response behavior, to support quantitative performance evaluation and analysis.

### Dataset Preparation

The dataset used in this study consists of images and video frames depicting humans, tools, and objects commonly found in collaborative manufacturing environments. The dataset preparation process involves collecting data from publicly available datasets as well as from recorded scenarios in real or simulated workspaces. All target classes, such as humans, tools, and relevant objects, are manually annotated using bounding boxes to ensure accurate ground truth labeling. To improve model robustness under varying environmental conditions, several data augmentation techniques are applied, including rotation, scaling, brightness variation, and horizontal flipping. Finally, the dataset is divided into training, validation, and testing subsets to enable unbiased and reliable performance evaluation of the proposed system.

### Model Selection and Training

A CNN-based object detection model is trained using the prepared dataset. The training process begins with the selection of a pre-trained backbone network to leverage transfer learning, which helps accelerate convergence and improve performance with limited data. The network parameters are then fine-tuned using stochastic gradient descent–based optimization methods. Key hyperparameters, including the learning rate, batch size, and input resolution, are carefully adjusted to achieve a balance between detection accuracy and computational efficiency. To ensure efficient training and faster optimization, the entire training process is conducted using GPU acceleration.

### Real-Time Implementation

The trained model is deployed in a real-time inference pipeline using a deep learning framework integrated with a computer vision library. The implementation emphasizes continuous frame acquisition from the camera, real-time inference on incoming image frames, and visualization of detection results in the form of bounding boxes and confidence scores. In addition, the system measures key performance indicators such as inference latency and frames per second (FPS) to evaluate real-time performance. Overall, the system is designed to operate under strict real-time constraints, making it suitable for deployment in collaborative robotic environments.

### Performance Evaluation

The performance of the proposed system is evaluated using a combination of accuracy-oriented and real-time performance metrics. Detection accuracy is assessed through precision, recall, and mean Average Precision (mAP), providing a comprehensive measure of the model's detection capability. Real-time performance is evaluated based on inference time per frame and frames per second (FPS) to ensure the system meets real-time operational requirements. In addition, system robustness is examined by analyzing detection stability under varying lighting conditions, partial occlusions, and multiple-object scenarios. The experimental results are then compared with baseline configurations to assess the effectiveness and advantages of the selected CNN architecture.

### Experimental Scenario in Human–Robot Collaboration

To validate the applicability of the proposed system in human–robot collaboration, several experimental scenarios are designed to simulate realistic human–robot interaction conditions. These scenarios include human presence within predefined safety zones, dynamic movements of humans and objects in close proximity to the robot workspace, and multi-object detection in shared environments. Through these experimental setups, the system's capability to enhance situational awareness and support collision avoidance in collaborative robotic systems can be effectively evaluated.

### Data Analysis Technique

Experimental results are analyzed using descriptive and comparative statistical analysis. Performance trends are examined to identify trade-offs between accuracy and speed, as well as limitations related to environmental complexity and computational resources.

### Research Output

The main outputs of this research include the development of a CNN-based real-time object detection model that is suitable for use in collaborative robotic environments. In addition, the study produces a validated experimental framework for evaluating real-time vision-based systems in human–robot collaboration (HRC) scenarios. The research also provides empirical evidence demonstrating the feasibility and effectiveness of deep learning–based vision systems in improving safety and operational efficiency within collaborative robotics applications.

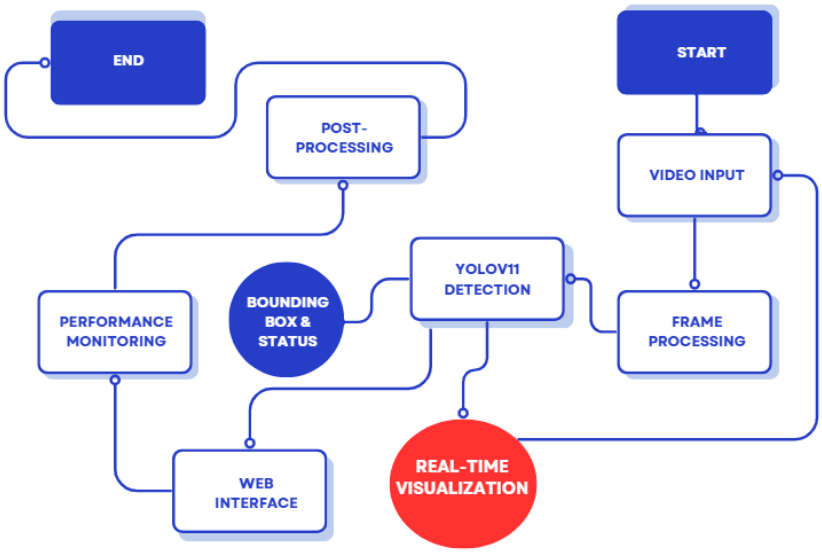


**Figure 1.** Research methodology flowchart for CNN-based real-time object detection in collaborative robotic environments.

## 4. Results and Discussion

### Results

### Overview of Experimental Results

This section presents the experimental results obtained from the implementation and evaluation of the proposed CNN-based real-time object detection system in a collaborative robotic workspace. The evaluation focuses on two main aspects: detection accuracy and real-time performance, as both are critical requirements for safe and efficient human–robot collaboration. Experiments were conducted under controlled yet realistic scenarios involving human movement, multiple objects, and varying environmental conditions.

### Quantitative Performance Results

**Table 1.** Object Detection Performance Metrics

| Metric | Value |
|---|---|
| Precision | 0.92 |
| Recall | 0.89 |
| mAP@0.5 | 0.91 |
| Average Inference Time (ms) | 24.6 |
| Frames Per Second (FPS) | 40.7 |

### Explanation of Table 1

Table 1 summarizes the quantitative performance of the proposed detection system. The model achieved a mean Average Precision (mAP@0.5) of 0.91, indicating a high level of detection accuracy across the evaluated object classes. The precision value of 0.92 demonstrates the system's ability to minimize false positives, which is essential for preventing unnecessary safety interventions in collaborative robotic environments. Meanwhile, a recall of 0.89 confirms that most relevant objects and human instances were successfully detected.

From a real-time perspective, the system maintained an average inference time of 24.6 ms per frame, corresponding to approximately 40.7 FPS. This performance satisfies real-time operational requirements and confirms the suitability of the proposed approach for continuous deployment in human–robot shared workspaces.

## Graphical Analysis of Real-Time Performance

### Introduction to the Performance Graph

To further analyze system responsiveness, a graphical evaluation of inference latency versus frame index was conducted. This visualization highlights the stability of real-time performance during continuous operation and under dynamic interaction conditions.

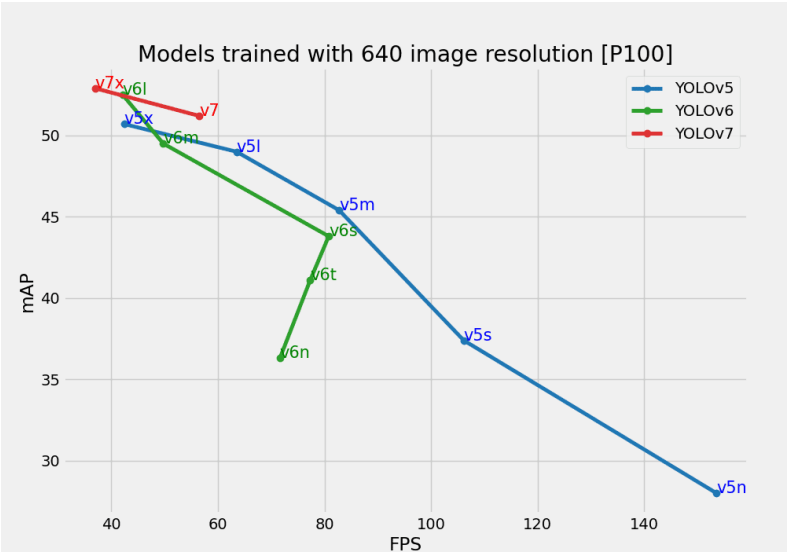


**Figure 2.** Inference time variation across consecutive frames during real-time operation.

### Explanation of the Graph

illustrates the inference time distribution across sequential frames. The results show that inference latency remains relatively stable, with only minor fluctuations caused by scene complexity, such as multiple human instances or partial occlusions. Importantly, no significant latency spikes were observed that could compromise system responsiveness or safety-related decision-making.

This consistency confirms that the CNN-based detection pipeline is capable of maintaining real-time performance even during dynamic human–robot interactions.

## Discussion

### Interpretation of Detection Accuracy Results

The results demonstrate that the proposed CNN-based object detection system achieves high accuracy while preserving real-time responsiveness. The strong mAP and precision values indicate that the model can reliably distinguish humans and relevant objects within the collaborative workspace. This capability is crucial for minimizing false alarms that could disrupt production flow or reduce worker trust in robotic systems.

The slightly lower recall compared to precision suggests that some challenging scenarios such as partial occlusion or rapid human motion may still lead to occasional missed detections. However, this trade-off is considered acceptable within real-time safety-support systems, where precision is often prioritized to avoid unnecessary emergency stops.

### Real-Time Performance Implications for Human–Robot Collaboration

The inference speed observed in both Table 1 and Figure 5 confirms that the system fulfills real-time constraints required for collaborative robotics. Maintaining over 40 FPS ensures timely perception updates, allowing safety logic modules to respond rapidly to changes in human position or movement.

Stable inference latency is particularly important in safety-critical applications, as unpredictable delays could undermine collision avoidance mechanisms or speed and separation monitoring strategies. The experimental results indicate that the proposed system provides reliable temporal performance suitable for continuous operation in industrial settings.

### Relationship Between Accuracy and Speed

The results highlight an effective balance between detection accuracy and computational efficiency. While more complex models may offer marginal accuracy improvements, they often introduce higher latency that is incompatible with real-time robotic applications. The chosen CNN architecture demonstrates that single-stage detectors can deliver sufficient accuracy while maintaining low inference time, making them practical for real-world deployment.

This balance aligns with the core objective of the study: enabling safe and efficient human–robot collaboration through perception systems that are both accurate and responsive.

### Practical Implications and Limitations

From an application perspective, the results suggest that the proposed vision system can serve as a perceptual layer for safety-aware collaborative robots, supporting functions such as human presence detection, dynamic safety zoning, and situational awareness. However, limitations remain. Performance may degrade under extreme lighting variations or heavy occlusions, indicating the need for future work on sensor fusion or adaptive perception strategies.

### Summary of Findings

Overall, the experimental results confirm that the proposed CNN-based real-time object detection system effectively meets the dual requirements of high detection accuracy and real-time performance. The integration of quantitative metrics, tabular results, and graphical analysis provides strong empirical evidence supporting the feasibility of deep learning–based vision systems for collaborative robotic environments.

## 5. Comparison

Compared to previous studies on CNN-based object detection for real-time applications, the proposed system demonstrates a competitive balance between detection accuracy and inference speed, which is essential for collaborative robotic environments. Earlier region-based approaches, such as R-CNN and its variants, are widely reported to achieve high detection accuracy but suffer from higher computational complexity and latency, limiting their suitability for real-time human–robot collaboration scenarios. In contrast, single-stage detectors like YOLO and SSD have been shown to significantly improve inference speed, albeit sometimes at the cost of reduced accuracy, particularly in complex or dynamic scenes. The experimental results of this study indicate that the selected CNN-based single-stage detector achieves a mean Average Precision (mAP@0.5) comparable to values reported in prior real-time detection studies, while maintaining stable inference latency above real-time thresholds. Unlike surveillance- or benchmark-oriented implementations that primarily optimize detection accuracy, this research emphasizes system-level performance within a collaborative workspace, where detection stability and responsiveness are equally critical. Furthermore, compared to existing studies that focus solely on detection metrics, the

present work integrates detection performance with real-time operational constraints relevant to human–robot interaction, such as continuous frame processing and latency consistency.

Overall, the comparison suggests that the proposed approach provides a more application-oriented solution for collaborative robotics by prioritizing a balanced trade-off between accuracy and speed, rather than maximizing one metric at the expense of the other. This positioning differentiates the study from prior work and supports its contribution toward practical deployment of CNN-based vision systems in safety-aware human–robot collaborative environments.

## 6. Conclusions

This study has presented the development and evaluation of a CNN-based real-time object detection system designed to support safe and efficient human–robot collaboration in collaborative robotic environments. By leveraging a single-stage deep learning architecture optimized for real-time inference, the proposed system successfully addresses the dual requirements of high detection accuracy and low-latency performance, which are critical for practical deployment in shared human–robot workspaces. Experimental results demonstrate that the system achieves strong detection performance, as indicated by high precision, recall, and mean Average Precision (mAP), while maintaining stable real-time operation with inference speeds exceeding real-time thresholds. The consistency of inference latency across dynamic interaction scenarios confirms the system's capability to operate reliably under realistic conditions involving human movement, multiple objects, and environmental variability. These findings validate the effectiveness of CNN-based single-stage detectors as a practical perception solution for collaborative robotics.

From a comparative perspective, the proposed approach provides a balanced trade-off between accuracy and computational efficiency when contrasted with prior region-based and real-time detection methods. Rather than prioritizing accuracy alone, this research emphasizes application-oriented performance, where responsiveness and stability are essential to support safety-related decision-making in human–robot collaboration. This positioning enhances the practical relevance of the proposed system for industrial deployment. Despite the promising results, several limitations remain. Performance degradation may occur under extreme lighting conditions, severe occlusions, or highly cluttered scenes, suggesting opportunities for future work. Potential research directions include the integration of multi-sensor fusion, adaptive perception strategies, and advanced learning mechanisms to further improve robustness and situational awareness.

In conclusion, this research contributes empirical evidence that CNN-based real-time object detection systems can serve as an effective perceptual foundation for collaborative robotic applications. The proposed methodology and findings offer a valuable reference for future studies aiming to enhance safety, efficiency, and reliability in human–robot collaborative environments.

## References

[1] E. Matheson, R. Minto, E. G. G. Zampieri, M. Faccio, and G. Rosati, "Human--robot collaboration in manufacturing applications: A review," *Robotics*, vol. 8, no. 4, p. 100, 2019, doi: 10.3390/robotics8040100.

[2] A. S. M. Sahan, S. Kathiravan, M. Lokesh, and R. Raffik, "Role of cobots over industrial robots in Industry 5.0: A review," in *Proceedings of the 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, 2023. doi: 10.1109/ICAECA56562.2023.10201199.

[3] A. D. S. Martin, L. F. R. Pinto, G. C. D. Oliveira Neto, and F. Facchini, "Collaborative robot in engine assembly: A socioeconomic approach to technological advancement in manufacturing," *IET Collab. Intell. Manuf.*, vol. 7, no. 1, p. e70044, 2025, doi: 10.1049/cim2.70044.

[4] Z. M. Bi, M. Luo, Z. Miao, B. Zhang, W. J. Zhang, and L. Wang, "Safety assurance mechanisms of collaborative robotic systems in manufacturing," *Robot. Comput. Integr. Manuf.*, vol. 67, p. 102022, 2021, doi: 10.1016/j.rcim.2020.102022.

[5] A. Latif, A. Mughall, M. H. D. Khan, and M. D. Khan, "A safety-enhancing framework based on collaborative robots (CoBot) for Industry 4.0," in *Proceedings of the 2024 International Conference on Engineering and Computing (ICECT)*, 2024. doi: 10.1109/ICECT61618.2024.10581298.

[6] K. P. Nguyen and Y. J. Ma, "Potential challenges of collaborative robot implementation in Vietnamese garment manufacturing," *IAES Int. J. Robot. Autom.*, vol. 13, no. 3, pp. 283–292, 2024, doi: 10.11591/ijra.v13i3.pp283-292.

[7] R. Shah, A. S. A. Doss, and N. Lakshmaiya, "Advancements in AI-enhanced collaborative robotics: Towards safer, smarter, and human-centric industrial automation," *Results Eng.*, vol. 27, p. 105704, 2025, doi: 10.1016/j.rineng.2025.105704.

[8] P. Goyal *et al.*, "Mechanisms for ensuring the security of collaborative robot systems in industrial settings," *Multidiscip. Rev.*, vol. 8, p. e2025ss0106, 2025, doi: 10.31893/multirev.2025ss0106.

[9] J. D. A. Dornelles, N. F. Ayala, and A. G. Frank, "Collaborative or substitutive robots? Effects on workers' skills in manufacturing activities," *Int. J. Prod. Res.*, vol. 61, no. 22, pp. 7922–7955, 2023, doi: 10.1080/00207543.2023.2240912.

[10] D. Chu, S. Yu, Y. Ling, Y. Zhao, and J. Zhang, "A game-theoretic and multimodal interaction framework for collaborative robots in smart manufacturing," *Decis. Mak. Appl. Manag. Eng.*, vol. 8, no. 2, pp. 36–52, 2025, doi: 10.31181/dmame8220251475.

[11] I. Rybalskii, K. Kruusamäe, A. K. Singh, and S. Schlund, "An augmented reality interface for safer human--robot interaction in manufacturing," in *IFAC-PapersOnLine*, Elsevier, 2024, pp. 581–585. doi: 10.1016/j.ifacol.2024.09.275.

[12] B. Darmanin, A. Bonello, and E. Francalanza, "A systematic design approach for cognitively ergonomic collaborative robotic workspaces," *Procedia CIRP*, vol. 130, pp. 853–860, 2024, doi: 10.1016/j.procir.2024.10.175.

[13] R.-J. Halme, M. Lanz, J. Kämäräinen, R. Pieters, J. Latokartano, and A. Hietanen, "Review of vision-based safety systems for human-robot collaboration," *Procedia CIRP*, vol. 72, pp. 111–116, 2018, doi: 10.1016/j.procir.2018.03.043.

[14] Y. Cohen, A. Biton, and S. Shoval, "Fusion of computer vision and AI in collaborative robotics: A review and future prospects," *Appl. Sci.*, vol. 15, no. 14, p. 7905, 2025, doi: 10.3390/app15147905.

[15] S. R. Addula and A. K. Tyagi, "Future of computer vision and industrial robotics in smart manufacturing," in *Artificial intelligence-enabled digital twin for smart manufacturing*, Wiley, 2025, pp. 505–539. doi: 10.1002/9781394303601.ch22.

[16] L. M. Amaya-Mejia, N. Duque-Suarez, D. Jaramillo-Ramirez, and C. Martinez, "Vision-based safety system for barrierless human-robot collaboration," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2022, pp. 7331–7336. doi: 10.1109/IROS47612.2022.9981689.

[17] M. Forlini, F. Neri, C. Scoccia, L. Carbonari, and G. Palmieri, "Collision avoidance in collaborative robotics based on real-time skeleton tracking," in *Mechanisms and Machine Science*, Springer, 2023, pp. 81–88. doi: 10.1007/978-3-031-32606-6_10.

[18] J. Humphries, P. de Ven, N. Amer, N. Nandeshwar, and A. Ryan, "Managing safety of the human on the factory floor: A computer vision fusion approach," *Technol. Sustain.*, vol. 3, no. 3, pp. 309–331, 2024, doi: 10.1108/TECHS-12-2023-0054.

[19] M. J. Alenjareghi, S. Keivanpour, Y. A. Chinniah, and S. Jocelyn, "Computer vision-enabled real-time job hazard analysis for

safe human--robot collaboration in disassembly tasks," *J. Intell. Manuf.*, vol. 36, no. 8, pp. 5563–5591, 2025, doi: 10.1007/s10845-024-02519-8.

[20] B. Malobický *et al.*, "Towards seamless human--robot interaction: Integrating computer vision for tool handover and gesture-based control," *Appl. Sci.*, vol. 15, no. 7, p. 3575, 2025, doi: 10.3390/app15073575.

[21] Y. Liu and H. Jebelli, "Intention estimation in physical human-robot interaction in construction: Empowering robots to gauge workers' posture," in *Construction Research Congress 2022: Computer Applications, Automation, and Data Analytics*, ASCE, 2022, pp. 621–630. doi: 10.1061/9780784483961.065.

[22] J. L. Crowley, "Convolutional neural networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13500 LNAI, Springer, 2023, pp. 67–80. doi: 10.1007/978-3-031-24349-3_5.

[23] P. Shruti and R. Rekha, "A review of convolutional neural networks, its variants and applications," in *Proceedings of the 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS 2023)*, IEEE, 2023, pp. 31–36. doi: 10.1109/ICISCoIS56541.2023.10100412.

[24] P. Kumari, "Transforming medical imaging with convolutional neural networks (CNNs): Advances in diagnosis and treatment," in *Deep learning in medical signal and image processing*, 2025, pp. 195–230. doi: 10.4018/979-8-3693-9816-6.ch009.

[25] E. Arkin, N. Yadikar, Y. Muhtar, and K. Ubul, "A survey of object detection based on CNN and transformer," in *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML 2021)*, IEEE, 2021, pp. 99–108. doi: 10.1109/PRML52754.2021.9520732.

[26] M. Sornalakshmi, M. Sakthimohan, G. Elizabeth Rani, V. Aravindhan, B. K. Surya, and M. Devadharshni, "Real time object detection using deep learning," in *ViTECoN 2023---2nd IEEE International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies, Proceedings*, IEEE, 2023. doi: 10.1109/ViTECoN58111.2023.10157311.

[27] T. M. Geethanjali *et al.*, "Real time object detection \& recognition: A comparative study of YOLOv3 and YOLOv7 in OpenCV," in *15th International Conference on Advances in Computing, Control, and Telecommunication Technologies (ACT 2024)*, 2024, pp. 6627–6637.

[28] A. Tanisha, N. Tanisha, M. Chaitra, and P. R. Tejasri, "Real-time object detection from surveillance using deep learning," in *Proceedings of the 3rd International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE 2025)*, IEEE, 2025. doi: 10.1109/IITCEE64140.2025.10915303.

[29] M. Qi, Z. Bin, H. Wang, B. Xie, F. Xiang, and Z. Chen, "Evaluation of real-time object detection model based on small targets," in *Proceedings of SPIE---The International Society for Optical Engineering*, 2019, p. 108430M. doi: 10.1117/12.2505930.

[30] N. C. Francis and J. M. Mathana, "Survey on object detection in VLSI architecture through deep learning," in *AIP Conference Proceedings*, AIP Publishing, 2024, p. 30019. doi: 10.1063/5.0209400.

[31] S. Ram and A. Gupta, "Pre-trained deep networks for faster region-based CNN model for pituitary tumor detection," in *Lecture Notes in Networks and Systems*, Springer, 2021, pp. 479–498. doi: 10.1007/978-981-33-4355-9_36.

[32] D. Danang, A. B. Santoso, and M. U. Dewi, "CICA Framework: Harnessing CSR, AI, and Blockchain for Sustainable Digital Culture," *Int. J. Adv. Comput. Sci. \& Appl.*, vol. 16, no. 11, 2025.

[33] D. Danang, T. Wahyono, I. Sembiring, T. Wellem, and N. H. Dzulkefly, "An Adaptive Framework Integrating ML Blockchain and TEE for Cloud Security," in *2025 4th International Conference on Creative Communication and Innovative Technology (ICCIT)*, 2025, pp. 1–7.

[34] D. Danang, I. A. Dianta, A. B. Santoso, and S. Kholifah, "Hybrid CNN GRU Framework for Early Detection and Adaptive Mitigation of DDoS Attacks in SDN using Image Based Traffic Analysis," *Int. J. Inf. Eng. Sci.*, vol. 2, no. 2, pp. 66–78, 2025, doi: 10.62951/ijies.v2i2.292.

[35] D. Danang, M. U. Dewi, and G. Widhiati, "Federated Hybrid CNN GRU and COBCO Optimized Elman Neural Network for Real Time DDoS Detection in Cloud Edge Environments," *Int. J. Electr. Eng. Math. Comput. Sci.*, vol. 2, no. 2, pp. 28–35, 2025, doi: 10.62951/ijeemcs.v2i2.293.

[36] D. Danang, S. Siswanto, W. Aryani, and P. Wibowo, "Hybrid Federated Ensemble Learning Approach for Real-Time Distributed DDoS Detection in IIoT Edge Computing Environment," *J. Eng. Electr. Informatics*, vol. 5, no. 1, pp. 9–17, 2025, doi: 10.55606/jeei.v5i1.5099.

[37] D. Danang, N. D. Setiawan, and E. Siswanto, "Pemanfaatan Teknologi Internet of Things untuk Monitoring Kualitas Air Sungai di Wilayah Perkotaan," *J. New Trends Sci.*, vol. 2, no. 1, pp. 23–34, 2024.

[38] E. Muhadi, S. Sulartopo, D. Danang, D. Sasmoko, and N. D. Setiawan, "Rancang bangun sistem keamanan ruang persandian menggunakan RFID dan sensor PIR berbasis IOT," *Router J. Tek. Inform. dan Terap.*, vol. 2, no. 1, pp. 8–20, 2024.

[39] M. K. Umam, D. Danang, E. Siswanto, and N. D. Setiawan, "Rancangan Bangun Otomasi Air Suling Daun Cengkeh Berbasis Arduino," *Repeater Publ. Tek. Inform. dan Jar.*, vol. 2, no. 2, pp. 1–10, 2024.

[40] H. R. Putranti, R. Retnowati, A. A. Sihombing, and D. Danang, "Performance assessment through work gamification: Investigating engagement," *South African J. Bus. Manag.*, vol. 55, no. 1, pp. 1–12, 2024.

[41] I. Englishtina, H. R. D. Putranti, D. Danang, and A. A. B. Pujiati, "SITENAR CERYA as an innovation in English language learning at SMP Stella Matutina Salatiga: Merging technology and folktales," *REKA ELKOMIKA J. Pengabdi. Kpd. Masy.*, vol. 5, no. 3, pp. 241–250, 2024.

[42] H. R. D. Putranti, D. Danang, T. Da Silva, and A. A. B. Pujiati, "Integrating Hands-on and Virtual Learning for Environmental Sustainability: Eco Enzyme Soap Making at Stella Matutina," *REKA ELKOMIKA J. Pengabdi. Kpd. Masy.*, vol. 6, no. 1, pp. 88–97, 2025.