Benchmarking YOLOv8-YOLOv12 Models for Image-Space Near-Miss Detection in Construction Safety

Shahid Ali[1], Ammar Alzarrad[2], Sudipta Chowdhury[3], Husnu S. Narman[4]

[1]Department of Computer Sciences and Electrical Engineering, Marshall University, One John Marshall Drive, Huntington, WV 25755; e-mail: shahidali@marshall.edu

[2]Department of Civil Engineering, Marshall University, One John Marshall Drive, WV 25755; E-mail: alzarrad@marshall.edu

[3]Department of Mechanical and Industrial Engineering, Marshall University, One John Marshall Drive, Huntington, WV 25755; e-mail: chowdhurys@marshall.edu

[4]Department of Computer Sciences and Electrical Engineering, Marshall University, One John Marshall Drive, Huntington, WV 25755; e-mail: narman@marshall.edu

*** Correspondence:**
Shahid Ali
shahidali@marshall.edu

Keywords: Near-miss detection, construction safety monitoring, YOLO object detection, worker-equipment interaction, computer vision in construction.

Abstract

Automated near-miss detection can support proactive construction safety monitoring by identifying hazardous worker-equipment interactions before injuries occur. This study presents a systematic evaluation framework for comparing 24 publicly available pretrained YOLO object detection models, spanning YOLOv8 through YOLOv12, for image-based near-miss detection in construction environments. The evaluation used 38 publicly available construction-related videos, standardized to 1920 × 1080 resolution and sampled at 1 frame per second, resulting in approximately 903 evaluation frames across varied work-zone conditions, camera perspectives, object scales, clutter, motion blur, and partial occlusion. A rule-based pipeline identified person-equipment interactions using centroid-based image-space proximity under a primary 250-pixel threshold. Because complete camera calibration metadata were unavailable, this study treats this threshold as an operational image-space criterion rather than a universally calibrated physical distance. This study evaluates object-detection performance using precision, recall, and F1 score against a human-corrected consensus ground truth derived from YOLOv8x candidate annotations and independently reviewed by three domain experts. The analysis assesses near-miss performance using frame-level recall, spatial overlap, fuzzy centroid matching, and threshold-sensitivity analysis. Results showed that YOLOv8x, YOLOv8l, and YOLOv9c achieved the strongest near-miss performance. The findings highlight the importance of stable frame-level detection and provide an exploratory, reproducible benchmark of comparative YOLO model behavior under selected image-space construction-video conditions.

1 Introduction

Construction projects involve constantly changing work conditions, moving equipment, shifting site layouts, and frequent interactions between workers and machinery. These conditions make construction sites particularly vulnerable to safety risks, including near-miss incidents where workers narrowly avoid injury. According to the U.S. Bureau of Labor Statistics (BLS, 2026), there were 5,070 fatal work injuries in the United States in 2024, down 4.0% from 5,283 in 2023. The fatal work injury rate was 3.3 fatalities per 100,000 full-time equivalent workers in 2024, compared with 3.5 in 2023. These national figures highlight the continued need for proactive safety monitoring in high-risk industries such as construction, where proximity hazards and improper use of personal protective equipment remain persistent concerns (Olimat et al., 2025). Traditional safety monitoring, which often relies on manual inspections, is labor-intensive, subjective, and limited in its ability to detect hazards in real time (López et al., 2025; Lim et al., 2025). These limitations become more pronounced as project complexity and pace increase, particularly when brief worker-equipment interactions must be identified quickly (López et al., 2025).

Advances in computer vision and deep learning offer new opportunities for proactive safety monitoring and automated assessment of civil infrastructure. Recent studies have demonstrated the use of drones, depth sensors, and machine-learning models for roof-condition inspection, solar-panel site assessment, earthquake-damage assessment, railroad gauge measurement, railway bolt detection, and construction-site firearm/tool/person detection (Alzarrad et al., 2021; Alzarrad et al., 2022; Kizilay et al., 2024; Le et al., 2025; Alzarrad et al., 2025; Damai et al., 2025). The YOLO (You Only Look Once) family of object detectors is widely used in real-time applications because it balances detection speed and accuracy (Lim et al., 2025; Alotaibi and Ma, 2025). Prior studies have applied YOLO models to PPE compliance (López et al., 2025), behavior recognition (Alotaibi and Ma, 2025), drowsiness detection (Onososen et al., 2025), and risk-zone alerts (Dzeng et al., 2025). However, most existing studies focus on a single YOLO variant or a specific detection task, leaving limited evidence on how newer YOLO versions and model sizes perform in near-miss scenarios where spatial relationships between workers and equipment are central.

This study introduces a comparative framework to evaluate YOLOv8 through YOLOv12 for near-miss detection in construction environments. Using centroid-based proximity analysis, the proposed pipeline processes frame-level video data to identify worker-equipment interactions. Outputs include annotated images, event logs, and detection metrics such as F1 score, recall, and spatial overlap, allowing cross-version comparison of detection behavior and risk sensitivity. This study seeks to answer the following questions: (1) How do different YOLO versions and sizes compare in detecting near-misses? (2) How do YOLO model generation and model size relate to detection accuracy in spatial risk contexts? The study provides a reproducible evaluation framework for construction safety research by standardizing inputs and evaluation criteria. The remainder of this paper is organized as follows. The Literature Review summarizes prior YOLO-based safety-monitoring studies and identifies gaps in interaction-level near-miss evaluation. The Methodology section describes the video dataset, the model-selection process, the annotation protocol, the proximity-based near-miss logic, and the evaluation metrics. The Results section presents object-detection and near-miss performance across all 24 YOLO variants, including threshold-sensitivity analysis. The Discussion interprets the findings in relation to model size, detection consistency, deployment constraints, and false-positive/false-negative tradeoffs. The Conclusion summarizes the study's contributions, practical implications, and directions for future research.

2 Literature Review

This section reviews recent YOLO-based and computer vision approaches to construction safety monitoring, with emphasis on studies published within the last five years. The review focuses on three areas relevant to this study: object-level safety detection, interaction-level worker-equipment risk assessment, and methodological issues related to ground-truth validation and proximity-based near-miss definitions. Near-miss detection differs from conventional object detection because it requires evaluating spatial relationships between workers and equipment rather than identifying isolated objects alone. YOLO-based models play a central role in safety monitoring tasks such as PPE compliance, drowsiness detection, and collision alerts, with many studies reporting strong detection performance (Onososen et al., 2025; Mohy et al., 2025; Dzeng et al., 2025). However, these studies primarily address static PPE compliance, isolated unsafe behaviors, object-level detection, or single-model configurations. They rarely evaluate interaction-level risk, where the spatial relationship between workers and equipment determines whether a detection becomes safety-critical. Prior work highlights the impact of occlusion, background clutter, lighting variation, small-object detection, and edge-computing constraints on real-world deployment; however, these factors remain seldom examined within a standardized multi-model near-miss evaluation framework. Although prior research proposes proximity-based frameworks for near-miss detection (Lim et al., 2025), existing studies remain limited to specific domains or individual model configurations, and no study provides a systematic evaluation across multiple YOLO generations using real-world construction-related footage. Existing YOLO-based methods, therefore, still lack a standardized evaluation framework for interaction-level risk assessment in near-miss detection scenarios.

Over the past five years, computer-vision research has expanded the role of YOLO-based models in safety monitoring beyond static object recognition. Earlier video-surveillance studies demonstrated that YOLO detections can be integrated with tracking and data-association methods to improve robustness under scale variation, occlusion, and motion uncertainty (Ait Abdelali et al., 2021). Subsequent work has explored hybrid CNN-Transformer detectors, real-scene safety datasets, and data-augmentation strategies to address small-object detection, class imbalance, and environmental variability in complex scenes (Guo et al., 2022; Yongxiang et al., 2022; Jiang et al., 2024). In construction-specific applications, recent studies have applied YOLO-based models to PPE detection, unsafe-behavior recognition, site-risk assessment, and UAV-based monitoring, showing strong promise while also reporting persistent challenges related to lighting variation, occlusion, cross-site generalization, and edge-device deployment (Feng et al., 2024; Zou and Hu, 2024; Lyu et al., 2025; Zhou et al., 2025). Together, these studies show that the field has progressed from basic object detection toward more context-aware, deployment-oriented monitoring; however, most prior work still evaluates detection accuracy at the object or task level rather than assessing how model outputs translate into near-miss identification at the interaction level.

Beyond construction PPE and near-miss applications, related civil infrastructure studies further demonstrate the value of AI-enabled visual sensing for safety-critical assessment. Drone-based and neural network approaches support the evaluation of roof suitability for solar panel installation and the assessment of roof conditions, reducing reliance on manual inspection in elevated or difficult-to-access environments (Alzarrad et al., 2021; Alzarrad et al., 2022). Similarly, drone imagery and fine-tuned deep-learning models have been applied to earthquake-damage assessment, while depth-sensor and machine-learning approaches have supported railroad gauge measurement and missing-bolt detection for railway safety applications (Kizilay et al., 2024; Le et al., 2025; Damai et al., 2025). These studies reinforce the broader relevance of computer vision for automating visual inspection and safety monitoring. However, they do not directly address near-miss detection at the interaction level for worker-equipment interactions in dynamic construction scenes.

Concerns regarding ground-truth validity and evaluation bias further highlight limitations in existing studies. While advanced detection frameworks achieve strong performance in safety monitoring tasks (Lee et al., 2023; Alotaibi and Ma, 2025), most studies fail to report inter-annotator agreement, expert labeling qualifications, and mechanisms to mitigate circular validation in model-assisted annotations. A broader review of construction safety research indicates that most AI-driven monitoring studies omit transparent human validation protocols (Olimat et al., 2025). Moreover, prior reviews of occlusion-handling approaches show that object detectors often degrade under partial occlusion, cluttered backgrounds, and complex surveillance conditions, and these limitations are not fully captured by IoU-based metrics alone (Ouardirhi et al., 2024). These findings underscore the necessity of independently human-validated datasets and documented annotation reliability when evaluating near-miss detection in safety-critical environments.

Methodological inconsistency is also evident in the calibration of proximity thresholds used to define near-miss events. Prior studies acknowledge that pixel-based distance thresholds are highly sensitive to camera geometry, field of view, lens distortion, and mounting configuration, yet such parameters are rarely reported in sufficient detail (Lim et al., 2025; Maulana and Hardiansyah, 2025). The literature suggests that robust near-miss evaluation should combine sensitivity analysis across multiple thresholds with camera-aware calibration to establish meaningful correspondence between pixel distance and physical separation. Treating proximity thresholds as tunable hyperparameters aligns with best practices in spatial risk assessment, though explicit calibration documentation remains essential for reproducibility and cross-site generalization.

Addressing these gaps, the present study provides a comparative evaluation of 24 YOLO variants spanning YOLOv8-YOLOv12 for near-miss detection using publicly available construction-related video footage. By shifting the emphasis from isolated object-detection metrics to frame-level recall and interaction-level risk identification, this work establishes a reproducible benchmarking framework tailored to safety-critical applications. The findings demonstrate that consistent temporal detection and contextual spatial reasoning are more influential for near-miss identification than absolute proximity distance alone, offering practical insights beyond conventional object-detection evaluation paradigms. Future research should extend this framework through expanded human-validated datasets, camera-geometry-based calibration, temporal modeling, edge-deployment evaluation, and multimodal sensing to further enhance robustness under occlusion and dynamic site conditions. Table 1 summarizes representative studies from 2021 to 2025 across video surveillance, construction safety, UAV monitoring, PPE detection, occlusion handling, and proximity-based near-miss detection.

Table 1. Comparative Analysis of YOLO-Based Approaches for Interaction-Level Risk Assessment

Study	Domain	Method/Focus	Performance	Key Limitation
Ait Abdelali et al. (2021)	Video surveillance / traffic	YOLO + Kalman filtering + Hungarian tracking	94.10% detection accuracy; 92.50% tracking accuracy	Traffic domain; not construction near-miss detection

Guo et al. (2022)	Industrial object detection	MSFT-YOLO / hybrid YOLO-Transformer detection	Improved defect-detection performance	Industrial defect domain; transfer to construction safety untested
Yongxiang et al. (2022)	Real-scene object detection	Real-scene detection dataset for object detection	Dataset contribution	Limited safety semantics and no near-miss interaction labels
Lee et al. (2023)	Construction	YOLACT + DeepSORT + PPE	91.3% accuracy	No inter-annotator agreement reporting
Feng et al. (2024)	Construction safety	YOLO-based PPE / unsafe-behavior or site-risk detection	Reported improved construction-site safety detection	Limited interaction-level near-miss benchmarking
Jiang et al. (2024)	Construction / PPE detection	Data augmentation for robust PPE recognition	Improved robustness under limited data	Rare near-miss and worker–equipment interaction events underrepresented
Zou and Hu (2024)	UAV / small-object hazard detection	YOLO-based aerial small-object detection	Improved small-object detection	Precision–recall tradeoff under tiny targets and changing viewpoints
Onososen et al. (2025)	Construction	YOLOv8 + drowsiness detection	92% mAP	Single static behavior, not interactions
Alzarrad et al. (2025)	Construction safety	AI-driven detection of people, tools, and firearms on construction sites	Reported automated multi-class detection performance	Focused on object-level threat detection; not worker-equipment near-miss interaction benchmarking
Mohy et al. (2025)	Construction	Computer vision model for automated safety compliance	Reported automated safety-compliance detection performance	Focused on safety compliance; not interaction-level near-miss benchmarking

Dzeng et al. (2025)	Construction	YOLOv7 + dynamic collision alert	0.998 mAP	Simulated virtual environment, not real footage
Lim et al. (2025)	Traffic / proximity	YOLOv7CNeB + distance indicators	0.997 mAP	Single model, traffic domain only
Alotaibi and Ma (2025)	Construction	Vision Transformer + unsafe-worker behavior detection	93.2% precision	No ground-truth validation methodology
Ouardirhi et al. (2024)	Smart video surveillance / object detection	Survey of occlusion-handling approaches	Review study	Addresses occlusion in surveillance but not construction-specific near-miss spatial context
Lyu et al. (2025)	UAV / edge detection	Lightweight YOLO for aerial edge deployment	Reported improved edge-oriented detection	Cross-site domain shift and edge-performance tradeoffs remain
Zhou et al. (2025)	UAV / small-object detection	SMA-YOLO with multi-scale attention and feature fusion	Improved small-object detection performance	Small-object recall, occlusion, and deployment tradeoffs remain
Saeedizadeh et al. (2025)	Autonomous driving / image-based detection	Survey of deep-learning object-detection methods	Review of 90+ models	Deployment robustness under occlusion, low light, and complex scenes remains challenging
Olimat et al. (2025)	Construction safety / bibliometric review	Data-driven review of construction safety research	14,174 publications reviewed; 61.8% recent studies from 2016–2025	Most studies lack transparent validation

3 Methodology

This section describes the dataset, model selection process, near-miss detection logic, ground truth annotation protocol, and the metrics used to evaluate both object detection and interaction-level performance across 24 YOLO variants. This study introduces a systematic framework for evaluating the performance of state-of-the-art object detection models, YOLOv8 through YOLOv12, in identifying near-miss incidents on construction sites. The methodology encompasses frame-level video sampling, model-based object detection, rule-based proximity analysis for near-miss identification, and comparative evaluation using precision, recall, and spatial alignment metrics.

Video Dataset and Frame Sampling: The dataset comprised 38 publicly available construction-related videos retrieved from the Pexels construction-video search page. The exact source-video set is reported in Supplementary Table S1 using the downloaded Pexels video identifiers and filenames. The source videos included HD and UHD footage with original frame rates ranging from 24 to 60 fps and individual durations ranging from under 1 minute to approximately 2 minutes. Across the 38 videos, the combined duration was approximately 903.13 seconds, or 15 minutes and 3 seconds, yielding approximately 903 sampled frames at the 1 fps sampling rate. Because the source platform does not provide standardized research metadata describing scene type, camera placement, lighting, weather, occlusion, or worker-equipment interaction conditions, these characteristics were manually reviewed by the authors through visual inspection of the videos. This review confirmed that the video set included diverse construction and work-zone scenarios, including road resurfacing, coastal excavation, urban equipment operation, and manual equipment handling, with visible variation in camera angle, object scale, worker-equipment proximity, background clutter, motion blur, and partial occlusion. To standardize evaluation across videos and reduce temporal redundancy, all videos were processed at a normalized 1920 × 1080 resolution and sampled at one frame per second. Under this sampling strategy, each second of video contributed one evaluation frame. Each extracted frame served as a consistent input across all models.

Model Selection and Proximity-Based Near-Miss Detection: A total of 24 YOLO model variants, spanning five generations from YOLOv8 to YOLOv12, were evaluated. Each generation included configurations from lightweight nano (n) to extra-large (x). The Ultralytics interface handled model loading, and all models were evaluated using the same inference pipeline across the complete set of sampled frames. No model was fine-tuned on the study videos; all YOLO variants were evaluated using their publicly available pretrained weights. Therefore, the study compares pretrained model behavior under a controlled evaluation pipeline rather than comparing models retrained specifically for this dataset. To ensure consistency, each YOLO variant was applied to the same sampled frames, with the same normalized resolution, object classes, confidence settings, IoU matching criteria, centroid-distance calculations, and near-miss matching rules. No model-specific threshold tuning, preprocessing adjustment, or post-processing adjustment was applied. Therefore, differences in near-miss performance reflect differences in model outputs under a controlled evaluation pipeline rather than differences in preprocessing, post-processing, or evaluation criteria.

All models were evaluated under a shared inference configuration. Object detections were retained using a fixed confidence threshold of 0.50, and non-maximum suppression was applied using an IoU threshold of 0.45 across all YOLO variants before computing precision, recall, and F1 score. The detector output was restricted to safety-relevant classes of interest: person, helmet, vest, truck, excavator, crane, vehicle, machinery, and backhoe. Near-miss events were defined only between person detections and equipment detections belonging to the subset {truck, excavator, crane, vehicle, machinery, backhoe}. For each retained bounding box $b = (x_1, y_1, x_2, y_2)$, the centroid coordinates were computed as:

$$c_x = \frac{x_1 + x_2}{2}$$

$$c_y = \frac{y_1 + y_2}{2}$$

The normalized centroid coordinates were computed as:

$$\hat{c}_x = \frac{c_x}{W}$$

$$\hat{c}_y = \frac{c_y}{H}$$

where $W = 1920$ and $H = 1080$ represent the standardized frame width and height, respectively.

The Euclidean image-space distance between a person centroid c_p and an equipment centroid c_e was computed as:

$$d(p,e) = \sqrt{(c_{x,p} - c_{x,e})^2 + (c_{y,p} - c_{y,e})^2}$$

A person-equipment pair was classified as a near-miss candidate when $d(p,e)$ was less than or equal to the selected proximity threshold. The primary analysis used a 250-pixel centroid-distance threshold, and the same computation was repeated at 150, 200, and 300 pixels to assess threshold sensitivity.

A 250-pixel threshold was selected as the primary image-space proximity threshold for defining near-miss events. Because the analyzed public construction footage did not include complete camera calibration metadata, including focal length, mounting height, viewing angle, lens distortion, and ground-plane reference measurements, the 250-pixel threshold should not be interpreted as a universally calibrated physical distance. Instead, it functions as an operational threshold for comparing model behavior under a fixed video resolution and camera perspective. For deployment in site-specific monitoring systems, this threshold should be recalibrated using known reference distances, camera calibration, homography-based ground-plane mapping, multi-camera geometry, or depth-aware sensing. Because centroid distance is measured on the two-dimensional image plane, the same physical worker-equipment separation can produce different pixel distances depending on camera height, viewing angle, object depth, object size, lens distortion, and perspective projection. Therefore, the proposed proximity rule should be interpreted as an image-space screening criterion for comparative benchmarking rather than a direct physical measure of near-miss severity.

Annotation Protocol and Performance Evaluation: The framework generates CSV-based logs of object detections and near-miss events for each model. These logs support computation of object detection (OD) metrics, including precision, recall, and F1 score, using an IoU threshold of ≥ 0.3. The IoU threshold of 0.3 was selected because the evaluation videos contained small objects, partial occlusion, motion blur, and variable camera perspectives, where stricter localization thresholds may penalize detections that remain sufficient for downstream centroid-based interaction screening. This threshold was applied uniformly across all models and was used for comparative object-detection evaluation rather than for claiming fine-grained localization accuracy. The analysis compares model outputs with a human-validated consensus ground truth. YOLOv8x detections provide preliminary candidate annotations to reduce manual labeling effort and do not serve as final labels. A human-in-

the-loop annotation strategy mitigates circular validation bias associated with model-assisted pre-labeling. Three domain experts, Ammar Alzarrad (Ph.D., M.ASCE, Construction Safety), Husnu S. Narman (Ph.D., Computer Vision), and Sudipta Chowdhury (Ph.D., Industrial Safety), independently review, correct, and supplement the annotations. The review process removes false-positive detections caused by background clutter, corrects bounding-box localization errors, adds missing detections absent from YOLOv8x outputs, and verifies person-equipment interaction labels used for near-miss evaluation. Consensus discussion resolves disagreements in object classification, bounding-box placement, and interaction boundaries until unanimous agreement is achieved for each contested frame.

Annotation guidelines were defined before expert review to ensure consistency across reviewers. A "person" annotation included any visible worker or pedestrian within the sampled construction-related frame. "Equipment" was operationally defined as visible construction or work-zone machinery and vehicles capable of producing a worker-equipment proximity hazard, including trucks, excavators, cranes, backhoes, and other mobile vehicles or machinery such as loaders, rollers, pavers, and forklifts when visible. When equipment types were not represented as explicit YOLO class names, they were assigned to the closest retained equipment superclass, such as "vehicle" or "machinery," for consistency with the model-output filtering rules. Static background objects, construction materials, signs, cones, and distant or fully occluded machinery not involved in a potential worker-equipment interaction were not labeled as near-miss equipment. A near-miss-positive interaction was defined as a person-equipment pair whose bounding-box centroids, after review, satisfied the selected image-space proximity rule. The final consensus benchmark contained approximately 903 sampled frames, including 18 near-miss-positive sampled frames. These positive frames contained 67 consensus person-equipment near-miss interaction instances under the 250-pixel image-space proximity rule. The frame-level positive count was used for frame-level recall, while the interaction-instance count was used for interaction-level near-miss precision, recall, and F1 evaluation. To reduce dependence on YOLOv8x proposals, expert review was conducted in multiple passes: first, candidate YOLOv8x detections were checked for false positives and localization errors; second, each full frame was visually scanned for missed persons, equipment, and worker-equipment pairs absent from the candidate annotations; third, centroid-based proximity labels were verified for eligible person-equipment pairs; and fourth, disagreements were resolved through consensus discussion until a final human-corrected reference label set was produced.

To quantify annotation reliability before consensus resolution, inter-rater agreement was computed using a stratified random sample of 200 frames, selected to represent different video conditions and interaction types. Annotators independently labeled frame-level near-miss presence according to the predefined proximity rule. Agreement among the three annotators was measured using Fleiss' kappa, yielding $\kappa = 0.87$, indicating strong agreement for a safety-critical visual labeling task. Pairwise agreement rates exceeded 90% across all annotator combinations. The finalized benchmark was therefore a human-corrected, consensus-resolved ground truth dataset rather than the raw YOLOv8x output. This consensus dataset served as the reference for all comparative evaluations, including YOLOv8x itself. Therefore, the analysis does not compare YOLOv8x against its raw predictions but evaluates it against a revised ground truth that removes false positives, corrects localization errors, and adds missing expert-identified detections. This procedure reduces circular validation bias and ensures that performance metrics reflect model behavior against an independently reviewed benchmark rather than artifacts of the initial pre-labeling model.

Near-miss (NM) performance was evaluated using centroid normalization and fuzzy matching to assess detection accuracy at both the instance and frame levels. Fuzzy matching between predicted and reference near-miss events was performed within each frame using greedy nearest-neighbor one-to-

one matching. A predicted person-equipment interaction was matched to a reference interaction when the corresponding centroid distance was within a 50-pixel tolerance, with ties resolved by selecting the smallest centroid distance first. Each reference interaction could be matched to at most one predicted interaction. Additional metrics included frame-level recall, average IoU within expanded person-equipment interaction zones, and mean proximity distance across all interactions. For the interaction-zone IoU calculation, each person and equipment bounding box was expanded outward by 50 pixels on all sides. The intersection-over-union was then computed between the expanded person box and the expanded equipment box using the standard intersection area divided by union area formula. If the union area was zero, the IoU was assigned a value of 0.0. Visualization of results used bar plots, line charts, and confusion matrices. The final evaluation also included an internal consistency check to verify that raw near-miss event logs aligned with the computed NM metrics. To clarify the unit of analysis for the statistical comparison, the ANOVA was not conducted using a single aggregate NM F1 score per model. Instead, a per-frame NM F1 score was computed for each of the 18 near-miss-positive sampled frames by comparing model-predicted person-equipment interactions with the consensus interactions in that frame using the fuzzy matching procedure described above. Each of the 24 YOLO variants therefore contributed 18 frame-level NM F1 records, yielding 432 model-frame records. Model variant was treated as the grouping factor in a one-way ANOVA, followed by Tukey-adjusted post hoc comparisons. Because these observations were sampled from video data and may retain temporal dependence within source videos, the ANOVA was interpreted as an exploratory within-benchmark comparison rather than as population-level statistical validation. Temporal clustering by source video and repeated evaluation of the same frames across models were not modeled as formal random effects. Therefore, no population-level inference about real construction-site performance is claimed from the ANOVA results. The primary interpretation, therefore, emphasizes descriptive model rankings, score differences, threshold-sensitivity patterns, and consistency checks.

4 Results

The evaluation framework computes object detection (OD) and near-miss (NM) performance metrics for each model. Within this benchmark, a near-miss candidate was operationally defined when the pixel distance between a detected person and a piece of equipment was less than or equal to 250 pixels. In the final consensus benchmark, 18 near-miss-positive sampled frames were identified, containing 67 consensus person-equipment near-miss interaction instances under the 250-pixel image-space proximity rule. Figure 1 illustrates sample near-miss detections from construction site footage, highlighting hazardous person-equipment interactions across diverse scenarios, including road resurfacing, coastal excavation, urban operations, and manual equipment handling. These examples capture varied public-video conditions, including differences in resolution, lighting, motion blur, and camera angles, illustrating the pipeline's comparative behavior under visually diverse image-space conditions rather than validating robustness across all real jobsite deployments.

YOLOv8x achieved the highest object-detection F1 score (0.980) and the highest near-miss F1 score (0.930). As shown in Table 2, YOLOv8x obtained NM precision = 0.819, NM recall = 0.879, frame-level recall = 1.000, average person-equipment interaction-zone IoU = 0.256, and detected near-miss-positive frames = 18. Other strong performers included YOLOv8l (OD F1 = 0.953, NM F1 = 0.879), YOLOv9c (NM F1 = 0.863), and YOLOv11x (NM F1 = 0.860). YOLOv10b and YOLOv12l also performed above the field median, each achieving an NM F1 score of 0.820. As shown in Figure 2, YOLOv8x and YOLOv8l led the near-miss F1 rankings, followed by selected YOLOv9 and YOLOv11 variants. The declining trend beyond the top models illustrates the challenge of maintaining detection consistency and spatial reasoning in compact architectures. Table 2 summarizes object-detection and near-miss metrics for all 24 YOLO variants, ranked by near-miss F1 score.

Table 2. Object detection and near-miss performance across 24 YOLO models

Model	Version	OD F1	NM Precision	NM Recall	NM F1	Frame-Level Recall	Avg IOU (Person–Equip Zone)	Detected NM Frames
yolov8x	YOLOv8	0.980	0.819	0.879	0.930	1.000	0.256	18
yolov8l	YOLOv8	0.953	0.858	0.848	0.879	0.944	0.280	17
yolov9c	YOLOv9	0.923	0.856	0.813	0.863	0.889	0.281	18
yolov11x	YOLOv11	0.927	0.830	0.874	0.860	0.833	0.281	16
yolov8s	YOLOv8	0.804	0.817	0.835	0.831	0.833	0.278	17
yolov8n	YOLOv8	0.741	0.827	0.820	0.831	0.056	0.173	11
yolov11s	YOLOv11	0.832	0.826	0.829	0.828	0.778	0.253	15
yolov9m	YOLOv9	0.929	0.819	0.823	0.822	0.889	0.268	17
yolov12l	YOLOv12	0.911	0.827	0.822	0.820	0.889	0.237	18
yolov9t	YOLOv9	0.720	0.826	0.814	0.820	0.167	0.185	9
yolov10n	YOLOv10	0.713	0.819	0.812	0.820	0.278	0.171	13
yolov10b	YOLOv10	0.920	0.814	0.803	0.820	0.944	0.269	18
yolov9e	YOLOv9	0.936	0.840	0.817	0.819	1.000	0.229	18
yolov12x	YOLOv12	0.942	0.803	0.824	0.817	1.000	0.252	18
yolov10m	YOLOv10	0.899	0.819	0.826	0.816	0.833	0.248	16
yolov10l	YOLOv10	0.916	0.815	0.823	0.816	0.944	0.252	18
yolov8m	YOLOv8	0.910	0.855	0.839	0.815	0.944	0.266	18
yolov11n	YOLOv11	0.733	0.817	0.818	0.815	0.056	0.159	7

yolov9s	YOLOv9	0.836	0.814	0.814	0.814	0.889	0.296	18
yolov11l	YOLOv11	0.900	0.816	0.821	0.813	0.833	0.217	16
yolov11m	YOLOv11	0.898	0.812	0.814	0.811	1.000	0.247	18
yolov10s	YOLOv10	0.843	0.822	0.813	0.811	0.944	0.276	18
yolov12n	YOLOv12	0.754	0.825	0.825	0.811	0.222	0.188	12
yolov12s	YOLOv12	0.836	0.825	0.814	0.810	0.944	0.293	18

OD F1 is the object detection F1 score against the human-validated consensus ground truth initialized from YOLOv8x candidate annotations. NM Precision, NM Recall, and NM F1 are computed for near-miss events under the 250-pixel proximity rule. Frame-level recall measures the proportion of frames with at least one correctly detected near-miss, Avg IOU is the mean intersection-over-union within person-equipment interaction zones, and Detected NM Frames is the number of sampled frames in which the model flagged at least one near-miss candidate. This frame-level count is distinct from the 67 consensus person-equipment near-miss interaction instances used for interaction-level precision, recall, and F1 evaluation.

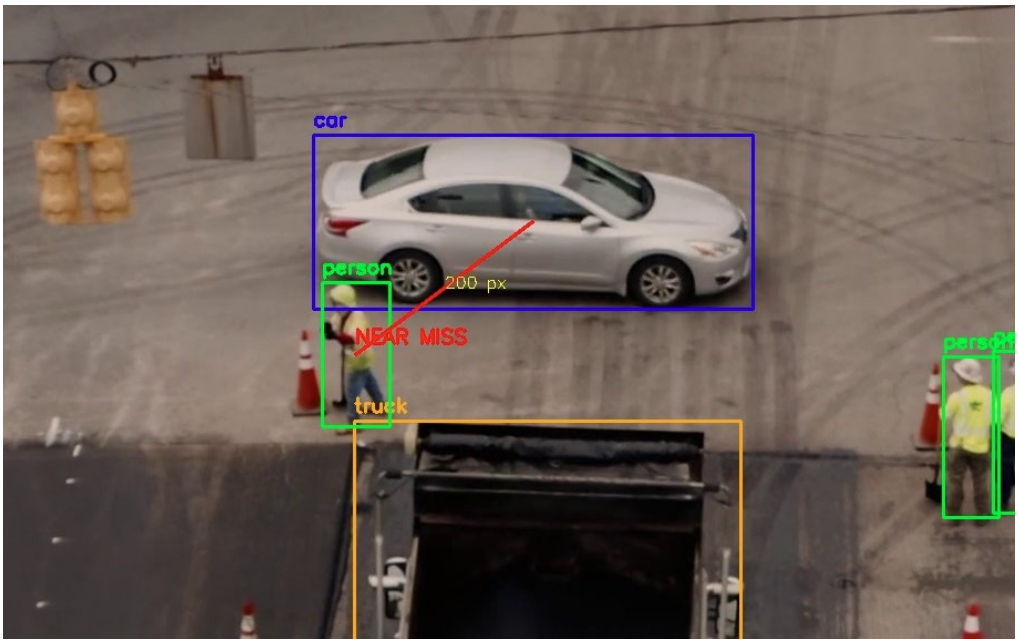

Figure 1: Annotated frames depicting detected near-miss interactions

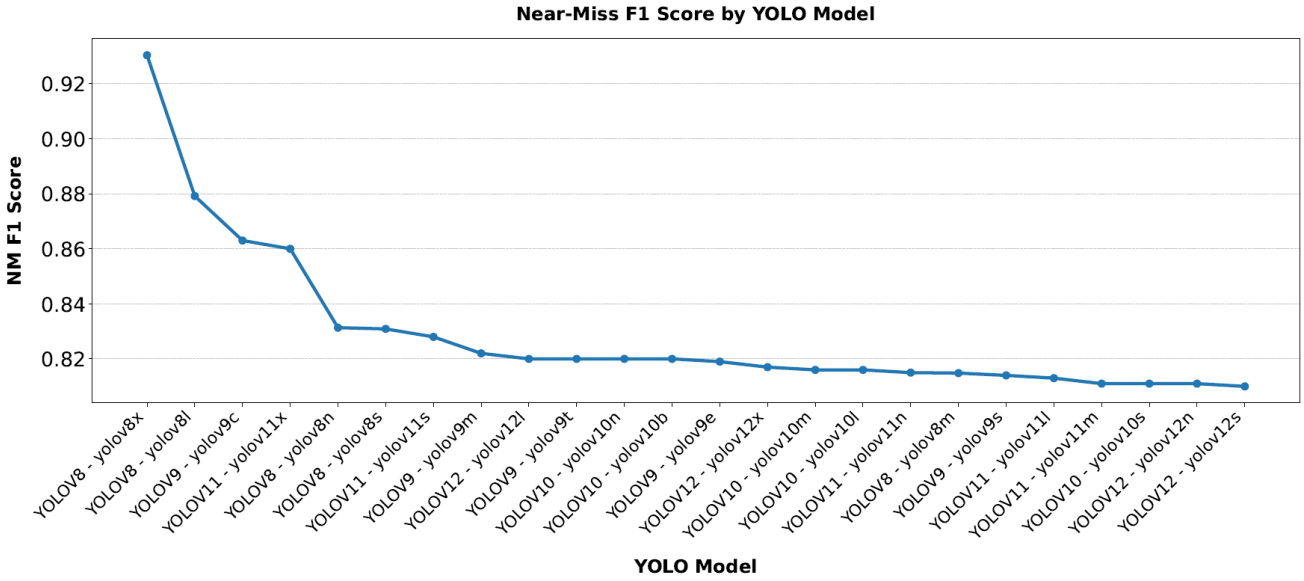

Figure 2: Near-Miss F1 Score by YOLO Model

Figure 3 illustrates differences in near-miss recall across YOLO variants. The leading models maintained stronger near-miss recall and more stable frame-level object coverage, suggesting that detection consistency is important for near-miss identification under cluttered and partially occluded construction-site conditions.

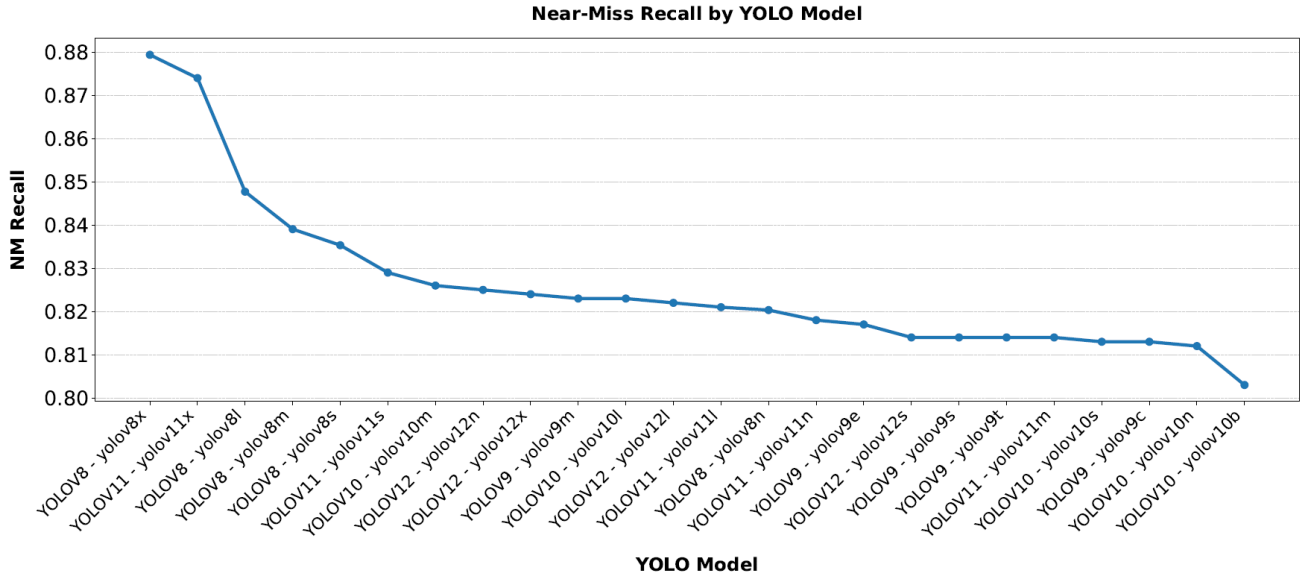

Figure 3: Near-Miss Recall by YOLO Model

As shown in Figure 4, near-miss precision was comparatively stable across several models, indicating that many variants produced spatially plausible near-miss detections when relevant objects were successfully detected. However, precision should be interpreted alongside frame-level recall because models with limited frame coverage may appear precise on the smaller subset of frames where detections occur, while still missing important near-miss events.

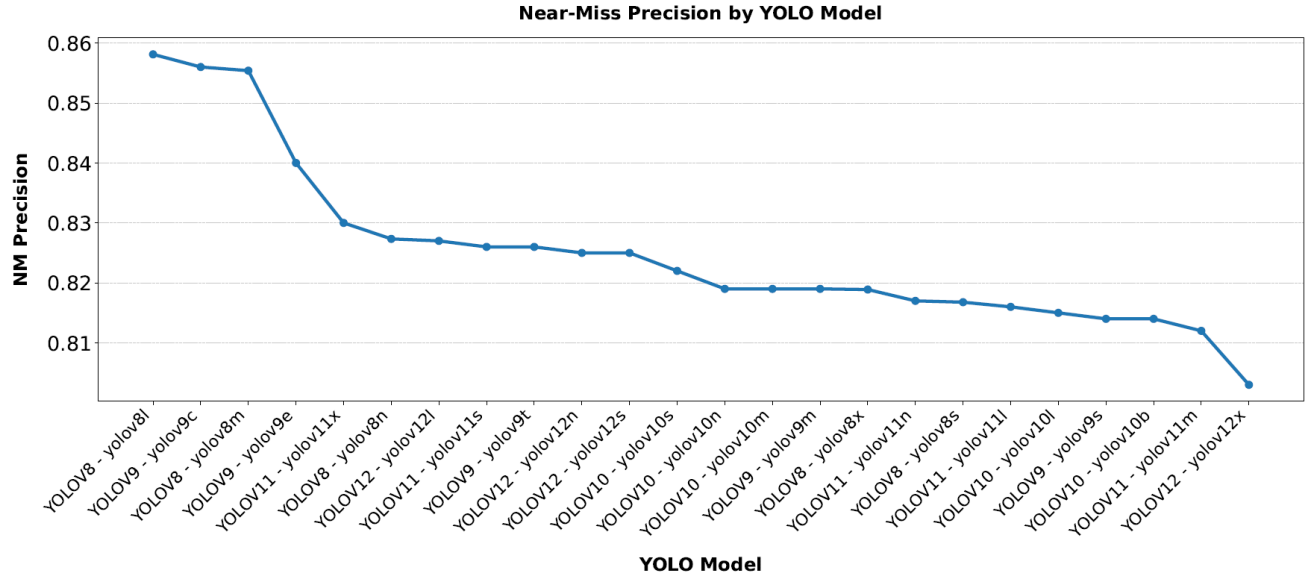

Figure 4: Near-Miss Precision by YOLO Model

A one-way ANOVA was conducted on observation-level NM F1 records rather than on a single aggregate NM F1 score per model. Each of the 24 YOLO variants contributed 18 near-miss-positive frame-level observations, yielding 432 model-observation records. The analysis showed significant variation in NM F1 across model variants, $F(23, 408) = 416.62$, $p < 0.001$. Tukey-adjusted post hoc

comparisons indicated that YOLOv8x remained separated from the lower-performing model cluster, defined as models with aggregate NM F1 scores at or below 0.820. These exploratory statistical results are consistent with the descriptive benchmark ranking, where YOLOv8x achieved the highest aggregate NM F1 score, followed by YOLOv8l, YOLOv9c, and YOLOv11x. However, because the observations were sampled from video footage and may retain temporal dependence within source videos, the ANOVA should be interpreted as exploratory evidence within the sampled benchmark rather than broad population-level statistical validation. Temporal clustering by source video and repeated evaluation of the same positive frames across models were not modeled as random effects, and the 18 positive frames represent a limited positive-event sample; therefore, the ANOVA results should not be interpreted as population-level evidence of real-world construction-site performance. Although YOLOv8x was used to generate the initial candidate annotations, its final performance was evaluated against the same human-corrected consensus ground truth used for all other models. This distinction is important because the final benchmark incorporated expert-added missed detections, corrected bounding boxes, and removed false positives rather than relying on unmodified YOLOv8x outputs.

Several lower-performing models, including YOLOv12s (NM F1 = 0.810), YOLOv12n (NM F1 = 0.811), YOLOv10s (NM F1 = 0.811), and YOLOv11m (NM F1 = 0.811), trailed the top-performing model by more than 0.10 NM F1 points. Models with particularly low frame-level recall, such as YOLOv8n (Frame-Level Recall = 0.056, NM F1 = 0.831) and YOLOv11n (Frame-Level Recall = 0.056, NM F1 = 0.815), correctly identified near-miss-positive frames less consistently. However, when these models produced detections, their proximity-based classifications remained broadly comparable to those of larger variants. YOLOv9t showed a similar pattern, with Frame-Level Recall = 0.167, NM F1 = 0.820, and only 9 detected near-miss-positive frames. This suggests that its lower near-miss coverage was primarily due to limited frame-level object detection rather than a failure of the proximity logic itself. Its object-detection F1 score was also relatively low at 0.720.

Correlation analysis showed a moderate positive relationship between NM F1 and frame-level recall ($r = 0.69$), compared with a weaker relationship between NM F1 and detected near-miss-positive frame count ($r = 0.35$). The correlation between NM F1 and average proximity distance was negligible ($r = -0.04$), suggesting that detection coverage and recall were more influential for near-miss performance than proximity distance alone. Internal consistency checks were conducted to verify that the frame-level detected near-miss counts and interaction-level NM precision, recall, and F1 scores were computed from the corresponding model event logs. Across all 24 models, OD F1 scores ranged from 0.713 to 0.980, NM F1 scores ranged from 0.810 to 0.930, frame-level recall ranged from 0.056 to 1.000, and average person-equipment interaction-zone IoU ranged from 0.159 to 0.296. As summarized in Table 3, model rankings remained broadly stable across proximity thresholds from 150 to 300 pixels, indicating that the comparative ranking was not driven solely by the selected 250-pixel cutoff. However, this threshold-sensitivity result does not remove the inherent limitations of centroid-based image-space proximity, because pixel distances remain affected by perspective, camera geometry, object depth, and scale variation. Accordingly, the reported results should be interpreted as comparative model performance rather than absolute physical distance measurements.

To assess sensitivity to the selected proximity threshold, the analysis recalculates near-miss F1 scores at 150, 200, 250, and 300 pixels for all 24 YOLO variants. Table 3 presents the results. The relative model-ranking pattern remains broadly stable across thresholds, indicating that the main findings do not depend solely on the 250-pixel cutoff.

Table 3. Threshold-Sensitivity Analysis of Near-miss F1 Scores Across Proximity Thresholds

Model	NM F1 @150px	NM F1 @200px	NM F1 @250px	NM F1 @300px
YOLOv8x	0.901	0.918	0.930	0.924
YOLOv8l	0.850	0.866	0.879	0.872
YOLOv9c	0.835	0.850	0.863	0.856
YOLOv11x	0.832	0.847	0.860	0.853
YOLOv8n	0.803	0.818	0.831	0.824
YOLOv8s	0.802	0.817	0.831	0.823
YOLOv11s	0.800	0.815	0.828	0.822
YOLOv9m	0.794	0.809	0.822	0.815
YOLOv12l	0.792	0.806	0.820	0.813
YOLOv9t	0.789	0.805	0.820	0.812
YOLOv10n	0.787	0.804	0.820	0.811
YOLOv10b	0.790	0.805	0.820	0.812
YOLOv9e	0.791	0.805	0.819	0.813
YOLOv12x	0.789	0.803	0.817	0.810
YOLOv10m	0.788	0.802	0.816	0.810
YOLOv10l	0.787	0.801	0.816	0.809
YOLOv11n	0.786	0.800	0.815	0.808
YOLOv8m	0.787	0.801	0.815	0.809
YOLOv9s	0.784	0.799	0.814	0.807
YOLOv11l	0.783	0.798	0.813	0.807
YOLOv11m	0.781	0.796	0.811	0.805
YOLOv10s	0.782	0.797	0.811	0.806

YOLOv12n	0.779	0.795	0.811	0.803
YOLOv12s	0.778	0.794	0.810	0.802

Figure 5 displays the number of sampled frames in which each YOLO model flagged at least one near-miss candidate. This frame-level count reflects coverage across the 18 consensus near-miss-positive sampled frames, but it is distinct from the 67 consensus person-equipment interaction instances used for interaction-level precision, recall, and F1 evaluation. Models with higher detected-frame counts generally provided broader frame-level coverage, while models with lower detected-frame counts may have missed near-miss-positive frames. Therefore, detected-frame count should be interpreted alongside precision, recall, F1 score, and frame-level recall rather than as a standalone indicator of model effectiveness.

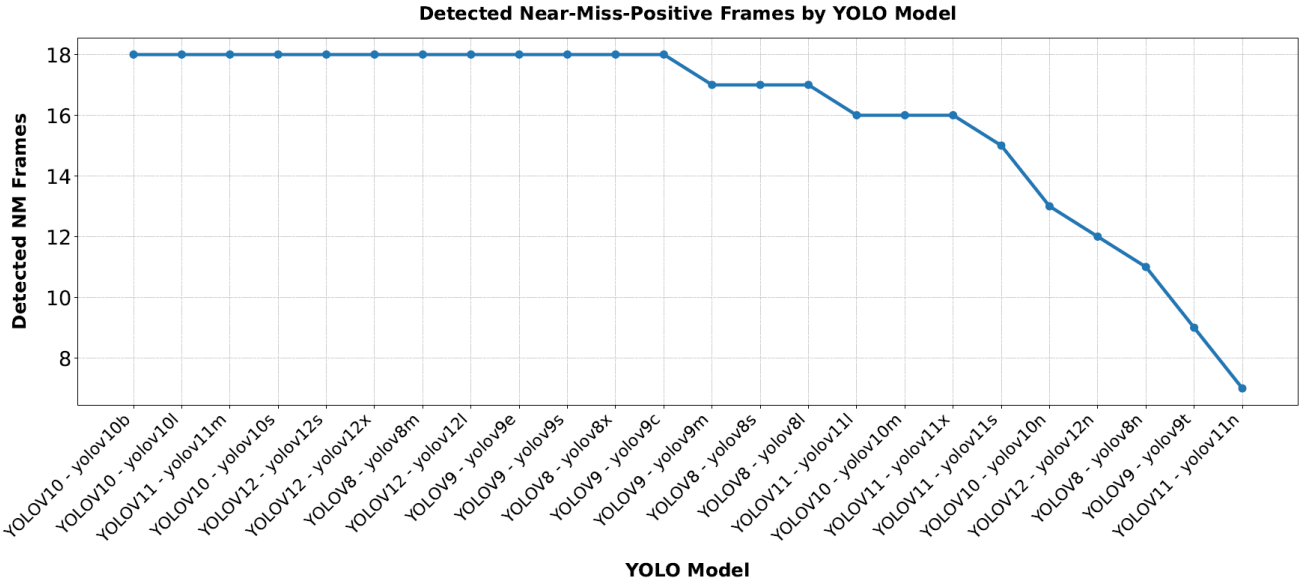

Figure 5: Detected Near-Miss-Positive Frames by Model Across YOLO Versions

5 Discussion

This study evaluated 24 YOLO model variants, spanning YOLOv8 through YOLOv12, for interaction-level near-miss detection in construction environments. The results show that stronger object-detection consistency and stable frame-level coverage are important for identifying hazardous worker-equipment interactions. Among the evaluated models, YOLOv8x achieved the highest near-miss performance (NM F1 = 0.930), followed by YOLOv8l (NM F1 = 0.879) and YOLOv9c (NM F1 = 0.863). These results indicate that larger or higher-capacity models generally performed better than lightweight variants in this task, particularly under visually complex conditions involving clutter, partial occlusion, motion blur, and varying object scales. These findings align with prior construction safety studies that have demonstrated the effectiveness of YOLO-based and deep-learning-based approaches for tasks such as drowsiness detection, PPE compliance, unsafe-behavior recognition, and safety-compliance monitoring (Onososen et al., 2025; Mohy et al., 2025; Alotaibi and Ma, 2025). However, the present study extends this body of work by focusing on interaction-level near-miss detection rather than isolated object or behavior recognition. In near-miss detection, accurate object localization alone is not sufficient; the model must also produce stable detections that support reliable spatial reasoning between

workers and equipment. This distinction is important because a missed person or equipment detection can directly affect whether a hazardous interaction is identified.

The performance of lightweight models further highlights the importance of frame-level coverage. For example, YOLOv8n and YOLOv11n both produced very low frame-level recall values of 0.056, while YOLOv9t achieved a frame-level recall of 0.167. These models flagged fewer near-miss events than higher-recall variants, suggesting that their limitations were primarily associated with inconsistent detection coverage rather than failure of the centroid-based proximity logic itself. This pattern is consistent with prior research showing that occlusion, background clutter, and complex surveillance conditions can reduce the reliability of object detection in real-world visual monitoring systems (Ouardirhi et al., 2024). For construction safety applications, these results suggest that model capacity should be considered not only in terms of object-detection accuracy but also in terms of whether the model can maintain reliable detections across frames. Correlation analysis provided additional insight into the factors influencing near-miss performance. Frame-level recall showed a moderate positive relationship with NM F1 performance ($r = 0.69$), while the relationship between NM F1 and average proximity distance was negligible ($r = -0.04$). This suggests that detection consistency and frame coverage were more influential for near-miss identification than the absolute closeness of detected objects alone. The threshold-sensitivity analysis also showed that model rankings remained broadly stable across proximity thresholds from 150 to 300 pixels. Therefore, the main comparative findings were not driven solely by the selected 250-pixel threshold. At the same time, because the proximity threshold was defined in image space rather than through full camera calibration, the results should be interpreted as comparative model performance rather than absolute physical distance measurements. Future deployment should incorporate camera calibration, ground-plane mapping, homography-based distance estimation, or depth-aware sensing to translate image-space proximity into physically meaningful worker-equipment distances. This limitation is especially important for construction scenes because workers and equipment may appear close in the image while being separated in depth, or may appear farther apart in pixels despite being physically close on the ground plane. As a result, centroid-based image-space proximity should not be interpreted as a complete near-miss definition or as a substitute for calibrated physical-distance estimation. Its role in this study is to provide a consistent operational rule for comparing detector behavior under the same image-space conditions.

The results also show that newer YOLO versions did not automatically outperform earlier variants. Although YOLOv10, YOLOv11, and YOLOv12 include architectural updates, their performance varied across model sizes. For example, YOLOv12l performed competitively, but YOLOv12n and YOLOv12s were among the lower-performing variants in near-miss F1. Similarly, YOLOv9t showed limited frame-level coverage despite belonging to a newer model generation. These results suggest that improvements in general object-detection architectures do not always translate directly into better performance for spatial proximity-based near-miss detection. Task-specific evaluation is therefore necessary when selecting models for construction safety monitoring, especially when the downstream goal involves interaction-level risk assessment rather than object detection alone. For practitioners, the findings suggest that model selection should depend on deployment context rather than accuracy alone. Within the present image-space benchmark, mid-to-large models such as YOLOv8x, YOLOv8l, and YOLOv9c achieved the strongest near-miss performance and may be preferable when the priority is reducing missed hazardous interactions. However, these models typically require greater computational resources, memory, and inference time than lightweight variants, which may limit their suitability for low-power edge devices, embedded cameras, or real-time on-site deployment. Smaller models may be more practical for resource-constrained environments, but their lower frame-level coverage may increase the risk of missed near-miss events. Therefore, deployment decisions should consider hardware availability, acceptable latency, camera density, network bandwidth, and whether

inference is performed locally at the edge or centrally on a server. However, this study did not directly measure inference latency, memory consumption, energy use, hardware-specific throughput, or real-time edge-device performance. Therefore, the deployment discussion should be interpreted as a practical consideration based on model size and detection behavior, not as a hardware-benchmarking result.

The tradeoff between false positives and false negatives is also central to practical use. False negatives are especially concerning in construction safety because they represent missed hazardous interactions and may prevent timely intervention. False positives, while generally less severe, can contribute to alarm fatigue, reduce worker trust, and increase the review burden for safety personnel. The results show that detected near-miss-positive frame count alone is not a sufficient indicator of model effectiveness. Models that flag more near-miss-positive frames may provide broader frame-level coverage but may also introduce additional false positives at the interaction level, while models with low frame-level recall may miss important hazardous frames. Accordingly, practical systems should tune detection and alert thresholds according to site-specific risk tolerance. High-risk zones involving heavy equipment may prioritize recall to reduce missed events, whereas lower-risk monitoring contexts may require stricter precision to avoid excessive alerts.

The external validity of the benchmark is limited by the use of publicly available Pexels videos rather than purpose-collected operational jobsite footage. Although the 38 videos varied in camera angle, object scale, worker-equipment proximity, clutter, motion blur, and partial occlusion, the dataset does not fully capture the frequency, severity, camera geometry, equipment movement patterns, weather conditions, site-specific work practices, or hazard distributions encountered in active construction monitoring. The absence of standardized metadata for camera placement, weather, lighting, and site operations also limits the extent to which the results can be generalized beyond the sampled video set. Therefore, the findings should be interpreted as comparative model behavior under selected image-space construction-video conditions, not as a broadly validated near-miss detection solution for field deployment. In addition, because the benchmark contained only 18 near-miss-positive sampled frames and 67 consensus near-miss interaction instances, the near-miss performance estimates may be sensitive to a small number of missed or additional detections and should be interpreted as exploratory rather than definitive. Future work should validate the framework using longer, purpose-collected, site-specific videos with documented camera geometry, calibrated distances, operational metadata, and safety-expert-labeled near-miss events.

Overall, this study demonstrates the value of evaluating construction safety models beyond conventional object-detection metrics. Near-miss detection requires stable object recognition, spatial reasoning, threshold sensitivity assessment, and careful interpretation of deployment constraints. The proposed framework provides a reproducible basis for comparing YOLO model variants in interaction-level safety monitoring. Future work should expand the dataset with longer and more diverse site-specific videos, incorporate calibrated physical-distance estimation, directly benchmark real-time latency, memory use, and throughput on edge and server hardware, and develop cost-sensitive metrics that separately weight false positives and false negatives according to construction safety priorities.

6 Conclusion

This research developed and applied an evaluation framework to assess 24 pretrained YOLO models, from YOLOv8 to YOLOv12, for construction near-miss detection. By combining object detection outputs with centroid-based proximity analysis, the framework examined how well different YOLO variants could identify potentially hazardous worker-equipment interactions in publicly available

construction-related video footage. The results indicate that near-miss detection depends heavily on consistent frame-level object coverage. YOLOv8x, YOLOv8l, and YOLOv9c produced the strongest overall performance, while several lightweight models showed reduced reliability in detecting near-miss events across frames. This suggests that model size and detection stability are important considerations when applying computer vision systems to interaction-level safety monitoring. The study also demonstrates the need to evaluate safety-monitoring models beyond standard object-detection metrics. Near-miss detection requires not only accurate object recognition, but also stable localization and reliable spatial interpretation of worker-equipment relationships. The proposed framework offers a reproducible basis for comparing YOLO-based safety systems. It can support future work on calibrated distance estimation, temporal tracking, direct latency and edge-device benchmarking, and cost-sensitive evaluation of false positives and false negatives.

7 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

8 Author Contributions

SA: Conceptualization, Visualization, Data curation, Methodology, Validation, Writing-original, draft, Writing-review and editing. AA: Methodology, Funding acquisition, Project administration, Writing-original draft. SC: Methodology, Software, Validation, Visualization, Writing-original draft. HN: Formal Analysis, Investigation, Writing-original draft.

9 Funding

The authors declare that financial support was not received for the research, authorship, and/or publication of this article.

10 References

Ait Abdelali, H., Derrouz, H., Zennayi, Y., Thami, R. O. H., and Bourzeix, F. 2021. “Multiple hypothesis detection and tracking using deep learning for video traffic surveillance.” *IEEE Access* 9, 164282-164291. https://doi.org/10.1109/ACCESS.2021.3133529.

Alotaibi, R. T. T., and Ma, S. 2025. “Real-time detection of unsafe worker behaviors via adaptive vision transformers in construction sites.” *Buildings* 15, 4205. https://doi.org/10.3390/buildings15224205.

Alzarrad, A., Emanuels, C., Imtiaz, M., and Akbar, H. 2021. “Automatic assessment of buildings’ location fitness for solar panel installation using drones and neural networks.” *CivilEng* 2, 1052-1064. https://doi.org/10.3390/civileng2040056.

Alzarrad, A., Awolusi, I., Hatamleh, M. T., and Terreno, S. 2022. “Automatic assessment of roof conditions using artificial intelligence and unmanned aerial vehicles.” *Front. Built Environ.* 8, 1026225. https://doi.org/10.3389/fbuil.2022.1026225.

Alzarrad, A., Ali, S., Chowdhury, S., and Narman, H. S. 2025. “Reducing gun-related incidents on construction sites: An AI-driven approach for automated detection of people, tools, and firearms.” In: *Computing in Civil Engineering 2025*. https://doi.org/10.1061/9780784486436.032.

Damai, A., Song, H., Narman, H. S., Lambert, A., and Alzarrad, A. 2025. "Enhancing railway safety: A machine learning approach for automated detection of missing track bolts." In: *Computing in Civil Engineering 2025*. https://doi.org/10.1061/9780784486436.024.

Dzeng, R.-J., Fan, B., and Hsieh, T.-L. 2025. "Dynamic collision alert system for collaboration of construction equipment and workers." *Buildings* 15, 110. https://doi.org/10.3390/buildings15010110.

Feng, R., Miao, Y., and Zheng, J. 2024. "A YOLO-based intelligent detection algorithm for risk assessment of construction sites." *J. Intell. Constr.* 2, 1-18. https://doi.org/10.26599/JIC.2024.9180037.

Guo, Z., Wang, C., Yang, G., Huang, Z., and Li, G. 2022. "MSFT-YOLO: Improved YOLOv5 based on transformer for detecting defects of steel surface." *Sensors* 22, 3467. https://doi.org/10.3390/s22093467.

Jiang, B., Zhu, Y., Zhou, Z., Zhang, Y., Xu, L., and Wang, J. 2024. "An improved safety belt detection algorithm for high-altitude work based on YOLOv8." *Electronics* 13, 850. https://doi.org/10.3390/electronics13050850.

Kizilay, F., Narman, M. R., Song, H., Narman, H. S., Cosgun, C., and Alzarrad, A. 2024. "Evaluating fine-tuned deep learning models for real-time earthquake damage assessment with drone-based images." *AI Civ. Eng.* 3, 15. https://doi.org/10.1007/s43503-024-00034-6.

Le, V. T., Song, H., Narman, H. S., Zhu, P., Alzarrad, A., Cisko, A., et al. 2025. "Automated measurement of horizontal gauge deviation in railroads using depth sensor camera and machine learning." *IEEE Access*. 13, pp. 215324-215338. https://doi.org/10.1109/ACCESS.2025.3641797.

Lee, Y.-R., Jung, S.-H., Kang, K.-S., Ryu, K.-C., and Ryu, H.-G. 2023. "Deep learning-based framework for monitoring wearing personal protective equipment on construction sites." *J. Comput. Des. Eng.* 10, 905-917. https://doi.org/10.1093/jcde/qwad019.

Lim, L. M., Yang, L., Zhu, W., Mohamed, A. S. A., and Ali, M. K. M. 2025. "Near miss detection using distancing monitoring and distance-based proximal indicators." *IEEE Access* 13, 48449-48468. https://doi.org/10.1109/ACCESS.2025.3548108.

López, L., Suárez-Ramírez, J., Alemán-Flores, M., and Monzón, N. 2025. "Automated PPE compliance monitoring in industrial environments using deep learning-based detection and pose estimation." *Autom. Constr.* 176, 106231. https://doi.org/10.1016/j.autcon.2025.106231.

Lyu, Y., Zhang, T., Li, X., Liu, A., and Shi, G. 2025. "LightUAV-YOLO: A lightweight object detection model for unmanned aerial vehicle image." *J. Supercomput.* 81, 105. https://doi.org/10.1007/s11227-024-06611-x.

Maulana, R. A., and Hardiansyah. 2025. "Implementation of YOLO algorithm for drowsiness detection as an additional safety feature in crane operation." *J. Inotera* 10, 113-120. https://doi.org/10.31572/inotera.Vol10.Iss1.2025.ID458.

Mohy, A. A., Bassioni, H. A., Elgendi, E. O., and Hassan, T. M. 2025. "Deep learning-enabled computer vision model for automated safety compliance in construction environments." *J. Inf. Technol. Constr.* 30, 1398-1430. https://doi.org/10.36680/j.itcon.2025.057.

Olimat, H., Alwashah, Z., Abudayyeh, O., and Liu, H. 2025. "Data-driven analysis of construction safety dynamics: Regulatory frameworks, evolutionary patterns, and technological innovations." *Buildings* 15, 1680. https://doi.org/10.3390/buildings15101680.

Onososen, A. O., Musonda, I., Onatayo, D., Saka, A. B., Adekunle, S. A., and Onatayo, E. 2025. "Drowsiness detection of construction workers using YOLOv8 and computer vision techniques." *Buildings* 15, 500. https://doi.org/10.3390/buildings15030500.

Ouardirhi, Z., Mahmoudi, S. A., and Zbakh, M. 2024. "Enhancing object detection in smart video surveillance: A survey of occlusion-handling approaches." *Electronics* 13, 541. https://doi.org/10.3390/electronics13030541.

Saeedizadeh, N., Jalali, S. M. J., Khan, B., and Mohamed, S. 2025. "Cutting-edge deep learning methods for image-based object detection in autonomous driving: In-depth survey." *Expert Syst.* 42, e70020. https://doi.org/10.1111/exsy.70020.

U.S. Bureau of Labor Statistics (BLS). 2026. "National Census of Fatal Occupational Injuries in 2024." News release, February 19. https://www.bls.gov/news.release/pdf/cfoi.pdf.

Yongxiang, G., Xingbin, L., and Xiaolin, Q. 2022. "YouTube-GDD: A challenging gun detection dataset with rich contextual information." *arXiv*. https://doi.org/10.48550/arXiv.2203.04129.

Zhou, S., Zhou, H., and Qian, L. 2025. "A multi-scale small object detection algorithm SMA-YOLO for UAV remote sensing images." *Sci. Rep.* 15, 9255. https://doi.org/10.1038/s41598-025-92344-7.

Zou, X., and Hu, Y. 2024. "Hidden danger detection and identification system of power transmission tower based on YOLOv11." *Acad. J. Sci. Technol.* 13, 224-231. https://doi.org/10.54097/rs28p954.

11 Data Availability Statement

The video data used in this study were obtained from publicly available construction-related videos on Pexels through the construction-video search page: https://www.pexels.com/search/videos/construction/. Because the Pexels search results page is dynamically updated and may be re-ranked over time, the exact benchmark source set is defined by the 38 Pexels video IDs and filenames listed in Supplementary Table S1. The frame-sampling procedure, benchmark composition, annotation protocol, consensus-review process, and shared inference/evaluation settings are described in the manuscript to support reproducibility. The authors do not redistribute the underlying videos; readers should access them through Pexels, subject to the platform's licensing and availability conditions. Additional implementation files or clarifications can be requested from the corresponding author.