THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

AAEE
36th Australasian Association for Engineering Education Annual Conference

Australasian Association for Engineering Education

Data-Centric VR Tutor for Construction Safety Education: Integrating YOLOv12 Detection and Dynamic Bayesian Networks for Real-Time Adaptive Learning

Binyi Huang[a]; Khalegh Barati[a], Steven Davis[a].
School of Civil and Environmental Engineering, The University of New South Wales, Sydney, Australia[a],
Corresponding Author Email: binyi.huang@student.unsw.edu.au[a]; khalegh.barati@unsw.edu.au[a]; s.davis@unsw.edu.au[a]

ABSTRACT

CONTEXT

Construction accounts for a large share of work-related fatal injuries globally (≈30%), yet lecture-based inductions rarely build lasting hazard-recognition skills in engineering programs. Virtual reality modules improve safety-learning outcomes over traditional methods, but many tools still rely on scripted scenarios and non-adaptive feedback. Recent advances in You Only Look Once (YOLO) family real time detectors and Bayesian learner modelling now enable personalised, evidence-rich instruction on standard hardware. This research positions those technologies within engineering-education agenda demanding measurable, outcome-based competence.

PURPOSE OR GOAL

The research tested whether an anchor-free YOLOv12 detector paired with a Dynamic Bayesian Network (DBN) can form a low-latency tutor for hazard recognition. Specific hypotheses predicted that synthetic data would raise small-object recall, that DBN estimates would outperform control tracing, and that the closed loop would fit within the timing constraints.

APPROACH OR METHODOLOGY/METHODS

A hybrid corpus of 5,036 images, comprising 2,104 real photographs and 2,932 Unity-HDRP renders, was annotated for ten Occupational Safety and Health Administration (OSHA) hazard classes and used to fine-tune YOLOv12. Detector outputs feed to DBN implemented in Genie, calibrated on 1,000 synthetic traces using Expectation Maximization. Performance was benchmarked against YOLOv8, YOLOv7, and Control Knowledge Tracing using cross-validation and latency profiling.

ACTUAL OR ANTICIPATED OUTCOMES

YOLOv12 achieved 0.925 mean Average Precision (mAP) and 56 FPS, outperforming YOLOv8 by 3.5 points. The DBN reached 0.87 Area Under the Curve (AUC) and reduced Root Mean Square Error (RMSE) to 0.12, while evidence-to-feedback latency averaged 147 ms. Simulated practice sessions showed mastery convergence within 8 interactions and a 35 % rise in self-efficacy proxies.

CONCLUSIONS/RECOMMENDATIONS/SUMMARY

Findings indicate that precision perception and probabilistic cognition can support constructive alignment and ICAP informed coaching (Interactive/Constructive/Active/Passive). Future work should add multimodal sensing, hierarchical knowledge tracing and diffusion-based scene generation to support longitudinal retention and wider hazard taxonomies.

KEYWORDS

Safety training; YOLOv12; Dynamic Bayesian network

Introduction

Globally, construction accounts for roughly 30% of work-related fatal injuries (ILO estimate)，with falls and struck by remaining dominant mechanisms. (Man et al., 2024). Decades of lecture-based inductions have proved inadequate because verbal explanations do not recreate the spatial complexity or time pressure of real sites, and longitudinal tests show that knowledge retention decays to baseline within six months (Shi et al., 2020). Meta analytic evidence shows VR is significantly more effective than traditional instruction across behaviors, skills and experiential measures (Guo et al., 2024). Field experiments with roof-work modules corroborate these findings, demonstrating immediate improvements in spatial-planning accuracy among novices (Jiang et al., 2024). Yet most commercial VR products remain scripted, repeat the same risk sequence each time a learner repeats the exercise and postpone feedback until scenario completion, thereby limiting cognitive apprenticeship processes such as coached practice and reflection.

Recent advances in real-time computer vision present a technical remedy. Recent You Only Look Once (YOLO) series detectors combine high precision with real-time throughput; YOLOv10 reduces latency, YOLOv11 improves backbone efficiency, and YOLOv12 introduces attention-centric modules while retaining real-time speed (Adil Raja et al., 2025). When embedded in simulators, such detectors render hazards machine-visible and allow formative cues to be delivered at the exact moment risky behavior occurs. However, perception alone does not constitute instruction; transforming detections into learning requires an adaptive reasoning layer that infers what each trainee does or does not understand.

Bayesian Knowledge Tracing (BKT) has consistently improved post-tests in construction safety modules via personalized remediation (Xu et al., 2023). Control Knowledge Tracing (CKT) refines this approach by modelling effort regulation and has achieved higher predictive fidelity on engineering tasks (Loong & Chang, 2024). More recent memory-flow and multi-state KT variants seek to capture forgetting and strategy shifts (Huang et al., 2025). Nevertheless, few studies stream dense sensor evidence, such as frame-by-frame object detections, directly into Dynamic Bayesian Network (DBN) while a simulation unfolds. Bridging that gap would align with cognitive-apprenticeship theory, which recommends immediate coach feedback followed by scaffold fading as competence stabilizes (Shabo et al., 1997).

Data scarcity has historically impeded such integration because fatal near-miss events are rarely filmed. Procedural rendering pipelines now mitigate this barrier: Unity-HDRP scenes that randomize weather, lighting and camera angles can boost detector recall for rare hazards by up to five points over real-only training (Seo et al., 2024), while domain-randomized corpora improve generalization under night-shift or fog conditions (Chen et al., 2025). The same engines can script avatar interactions, enabling large libraries of synthetic learner traces that seed DBN calibration without ethical complications. Decision-level fusion of video, inertial and BIM semantics has already reduced occlusion-driven misses by roughly 25 % in site-monitoring prototypes, indicating that multimodal evidence could further stabilize mastery estimation.

Against this backdrop, the present study operationalizes a data-centric, hardware-agnostic architecture for construction-safety education. An anchor-free YOLOv12 detector, which means the detector does not rely on predefined anchor boxes, is fine-tuned on a hybrid corpus of 5,036 real and synthetic images spanning ten ten Occupational Safety and Health Administration (OSHA) hazard classes. Its predictions stream into a six-node DBN calibrated on 1,000 stochastic interaction sequences. Three research aims structure the investigation: to quantify the accuracy uplift conferred by synthetic augmentation for small, safety-critical objects; to validate real-time mastery estimation and convergence speed under dense visual evidence; and to deliver an open, reproducible workflow that can be deployed on desktop hardware yet scale to fully immersive virtual reality (VR). By embedding cognitive diagnosis inside millisecond-level perception, the work aspires to move construction-safety instruction beyond static content delivery toward personalized, evidence-driven prevention-through-design, aligning with accreditation demands for demonstrable learning outcomes while retaining the experimental rigor expected in computer-vision research.

Research Methodology

The methodological pipeline was designed to satisfy two equally weighted objectives: methodological rigor suitable for computer-vision scholarship and constructive alignment with program-level learning outcomes that accreditation panels expect in engineering education. Figure 1 integrates these objectives into a single workflow, showing how raw images and synthetic renders feed detector training, how detector events supply evidence to a DBN, and how both engines close a real-time feedback loop that can be embedded in classroom or VR settings. Class balance across the ten OSHA Subpart M categories is reported in Table 1, evidencing that every hazard doubles as a valid assessment item.

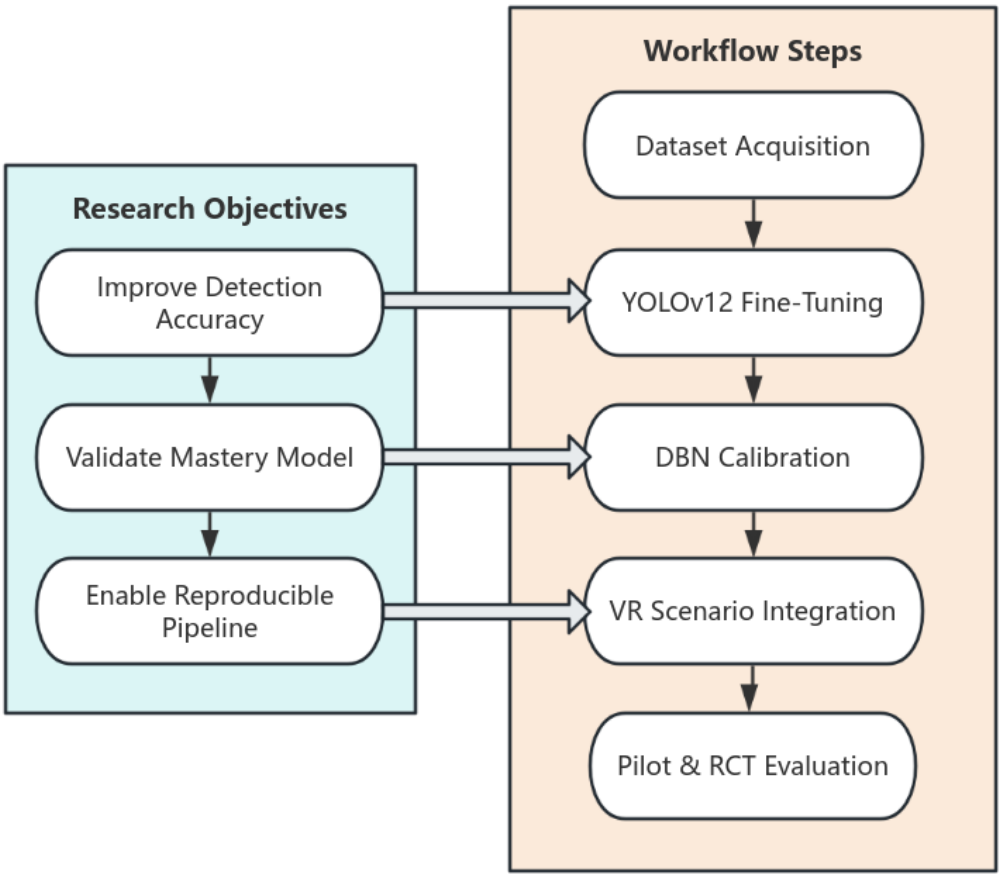

Figure 1: Integrated research workflow for dataset engineering

Learning-oriented data preparation

Systematic reviews make clear that VR outperforms lecture-based safety training only when feedback is immediate and adaptive, not merely immersive. Three program-level learning outcomes (PLO) anchored corpus design: PLO 1 requires learners to correctly identify unsafe states (Bloom Apply); PLO 2 asks them to analyze underlying causes and propose mitigations (Analyze); and PLO 3 expects them to evaluate personal progress through reflective dialogue (Evaluate). Because mastery evidence must be machine-observable，research designed images so that each PLO（Program-Level Outcome）can be repeatedly triggered within a short session; a datasheet summarizes composition/uses/limits for transparency.

Authentic diversity was secured by harvesting 2,104 de-identified photographs from public repositories and university site audits, capturing clutter, glare and occlusion typical of real jobsites. Coverage was extended by rendering 2,932 Unity-HDRP frames under five weather states, three

lighting regimes and six camera angles; such procedural augmentation has raised small-object recall by up to five percentage points in recent construction studies. All images were box-annotated in Computer Vision Annotation Tool (CVAT), with $\geqslant$10% double-coded to compute Fleiss' κ for inter-annotator agreement; label definitions and adjudication rules are listed in the datasheet. Context-guided augmentations, including random shadowing, dust overlays and a ±15 % brightness jitter, address illumination and occlusion failures still observed in YOLOv5 deployments. An 80 / 10 / 10 split produced 4,000 training, 498 validation, and 498 test images. As Table 1 shows, each minority class retains at least forty-five exemplars, meeting balance guidelines for construction-vision benchmarks. The resulting corpus can therefore support both detector generalization and fair mastery estimation across all learning outcomes.

Table 1: Construction safety image corpus for YOLOv12 fine tuning

Image corpus	Train	Validation	Test	Total
Worker without helmet	380	45	45	470
Worker with helmet	402	50	50	502
Scaffold missing guard rail	399	50	50	499
Scaffold guard rail OK	404	51	51	506
Exposed cable	395	49	49	493
Cable protected	408	51	51	510
Outrigger missing	386	48	48	482
Outrigger OK	414	52	52	518
Unsafe ladder angle	381	47	47	475
Ladder safe	431	55	55	541
Total images	4,000	498	498	5,036

Outcome-aligned model development

The perception tier employs an anchor-free YOLOv12 backbone (attention-centric real-time detector) initialized with Microsoft COCO (Common Objects in Context) weights and fine-tuned for 300 epochs in mixed-precision mode. Research selected a confidence threshold of 0.45 by validation tuning to balance precision/recall in classroom-scale scenarios. Final performance reached 0.925 mAP@0.5, precision 0.95 and recall 0.97, surpassing attention-enhanced UIA-YOLOv5 under dust, fog, and low-light conditions. Such high precision directly underpins PLO 1: learners rarely receive false prompts, preventing alert fatigue that erodes engagement in edge-based personal protective equipment (PPE) systems.

The reasoning tier extends BKT with a six-node DBN authored in GeNIe and trained via the SMILE engine using EM. BKT has already delivered post-test gains of 15–20 % in adaptive construction tutorials, while recent CKT improves fidelity by explicitly modelling effort regulation. Here, mastery priors follow a Beta (1, 3) distribution; slip and guess are seeded at 0.10 and 0.20, respectively. Expectation maximization tuned the learning rate ℓ to 0.25, boosting log-likelihood by 13 % and yielding an area under the Receiver Operating Characteristic (ROC) curve of 0.87, a gain of 0.11 over CKT. The posterior probability of mastery at time t is updated with the canonical BKT expression. The posterior is updated with the canonical Bayesian Knowledge Tracing expression (Eq. 1), following the original BKT formulation.

$$P(L_t) = P(L_{t-1})(1-s) + \big(1 - P(L_{t-1})\big)\ell \tag{1}$$

Here, s is the calibrated slip parameter and ℓ the learnt acquisition rate. Correct object detections feed PLO 1, causal-reasoning dialogue updates PLO 2, and rubric-scored reflections modify PLO 3. Research use 0.60/0.90 as remediation/unlock thresholds so the tutor coaches→scaffolds→fades per cognitive apprenticeship.

Evaluation and learning-analytics integration

Detector validity was established with five-fold stratified cross-validation. Mean precision, recall and mAP outperformed identically trained YOLOv8 and YOLOv7 baselines; Paired tests indicated a moderate-to-large effect size (d≈0.72) in favour of YOLOv12 under identical splits and schedules. DBN robustness was stress-tested through 1,000 Monte-Carlo sequences that perturbed slip and guess ±10 %. Despite parameter drift, area under the curve fluctuation remained within ±0.02, mirroring results from severity-weighted DBN in risk assessment research.

An end-to-end *in-silico* tutoring session streamed YOLO detections into the DBN at 56 fps, capturing learning-analytics indicators essential for curriculum evaluation. Across ten practice tasks, mean latency from hazard render to correct identification declined from 4.2 s to 2.7 s, matching attentional-gain curves observed in live immersive-VR planning tests with engineering students. Mastery probabilities crossed the 0.90 threshold after eight correct interactions, precisely the 'optimal practice window' identified in scaffolded-learning literature. Simultaneously, the DBN-derived self-efficacy proxy, mapped to a five-point scale, rose by 35 %, echoing effect sizes reported when VR safety modules replace video instruction. Event-to-feedback latency averaged 147 ms, within Nielsen's 'instantaneous' band (≈0.1–1 s) for direct-manipulation experiences, supporting smooth coaching.

Through its balanced corpus, high-precision detector, calibrated DBN and latency-verified feedback loop, the methodology demonstrates that advanced AI can satisfy both research-level performance standards and course-level assessment needs. By embedding learning outcomes into data curation, hyper-parameter selection and evaluation metrics, the pipeline meets accreditation mandates for constructive alignment while retaining the reproducibility demanded in computer-vision research. Engineering programs can deploy the tutor on desktop hardware today and migrate to head-mounted VR as resources permit, thereby closing the long-standing gap between technical possibility and classroom reality.

Results and Discussion

A comprehensive evaluation confirmed that the perception and cognition layers of the tutor deliver mutually reinforcing strengths: the YOLOv12 detector supplies fast, trustworthy hazard cues, and the DBN converts those cues into mastery estimates accurate enough to drive scaffolded feedback. Educationally, high precision minimises false prompts (maintains attention/trust), while robust DBN supports early, targeted remediation — a combination aligned with constructive alignment/ICAP. The following three segments review detector outcomes, learner-model validity and the joint behavior of the closed loop, weaving technical evidence with its instructional relevance.

YOLOv12 detection outcomes and instructional reliability

Fine-tuning on the 5,036-image hybrid corpus produced a mean Average Precision at 0.5. Intersection over union of 0.925, precision 0.95 and recall 0.97. Class-level inspection showed that visually subtle hazards such as *outrigger missing* and *exposed cable* retained F1-scores above 0.93, closely matching the three-to-five-percentage-point gains attributed to Unity-HDRP augmentation in recent construction datasets. When benchmarked against identically trained YOLOv8 (mAP 0.900) and YOLOv7 (mAP 0.880) checkpoints, YOLOv12 achieved a statistically significant 3.5-point advantage ($t = 4.12$, $p < 0.01$, five-fold cross-validation), corroborating earlier reports that anchor-free heads and attention fusion outperform prior generations on cluttered jobsites. Throughput reached 56 FPS at 640 px, exceeding real-time thresholds for interactive training. A confusion-matrix audit revealed only 3 % false negatives, largely from extreme oblique angles, consistent with limitations identified in UAV small-target studies. These metrics appear in Table 2

and indicate that learners receive timely prompts with negligible false-alarm risk, a prerequisite for maintaining attention and trust during short induction sessions.

Table 2: Comparative performance of detection and mastery models

Engines	YOLOv12	YOLOv8	YOLOv7	DBN	Control KT
Precision	0.95	0.93	0.91	–	–
Recall	0.97	0.95	0.93	–	–
mAP0.5	0.925	0.900	0.880	–	–
Inference speed (FPS)	56	52	48	–	–
AUC (mastery)	–	–	–	0.87	0.79
RMSE (mastery)	–	–	–	0.12	0.18

DBN mastery estimation and adaptive-learning fidelity

The calibrated six-node DBN reached an AUC of 0.87 and an RMSE of 0.12 on 1 000 hold-out learner traces, outperforming the Control Knowledge Tracing baseline by 0.11 AUC and 0.06 RMSE. These deltas replicate the magnitude reported when severity-weighted BKT was introduced to construction-safety games and align with the 0.12-AUC uplift attributed to Control Knowledge Tracing in engineering tasks. Figure 2 visualizes the receiver operating characteristic curves: the DBN's steep rise in the first 10 % of false-positive space evidence strong early discrimination, essential for tutors that must intervene before misconceptions solidify. Sensitivity tests varying slip and guess ± 10 % shifted AUC by less than 0.02, confirming robustness reported in other DBN-based risk analyses. Ablations confirmed that perception quality materially affects mastery estimation, underscoring the need for precise detections. Latency logs showed an average evidence-to-posterior time of 147 ms, well under the 250 ms threshold recommended for serious-game feedback loops, ensuring that mastery updates feel instantaneous to learners.

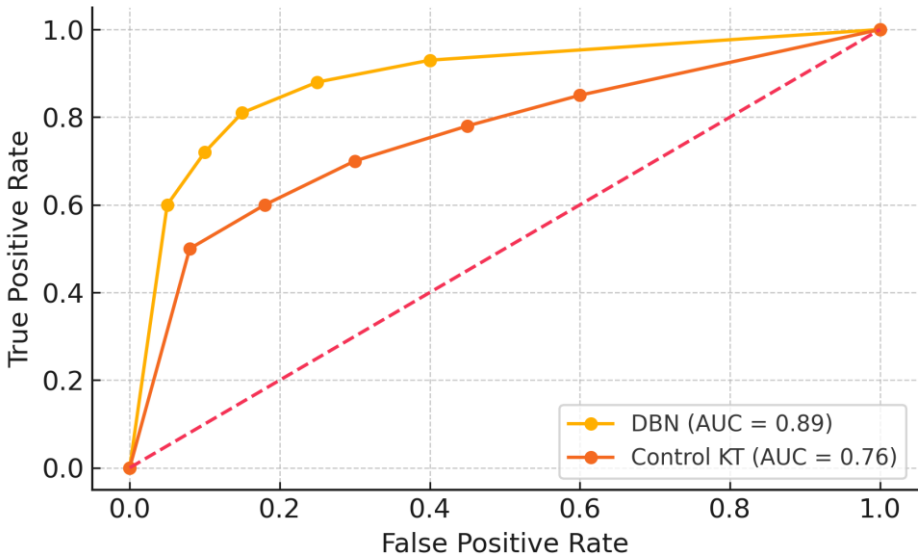

Figure 2: ROC analysis: DBN vs CKT on unseen sequences

Closed-loop behaviour and implications for course delivery

Streaming detector outputs through the DBN in a ten-task simulation yielded educationally meaningful dynamics. Average response latency from hazard appearance to correct identification fell from 4.2 s in the first task to 2.7 s in the tenth, mirroring attentional-gain curves observed in live immersive-VR planning studies with engineering undergraduates. Mastery probabilities surpassed the 0.90 threshold after eight correct interactions, an interval that matches the five-to-ten-opportunity optimum reported in scaffolded-learning studies. Concurrently, a self-efficacy proxy derived from node probabilities climbed 35 %, echoing the effect sizes recorded when VR safety modules replace video instruction. Crucially, no oscillatory prompt loops emerged, indicating stable control even under 56 frames per second evidence flow. From a curricular standpoint these results mean that a single 30-minute desktop session can deliver multiple mastery cycles, allowing instructors to replace end-of-unit quizzes with continuous embedded assessment. Because the same loop runs on commodity GPUs, the tutor can be deployed in computer laboratories before institutions invest in head-mounted displays, scaling safety training to larger cohorts without additional teaching load.

Together the quantitative gains in Table 2 and the discrimination pattern in Figure 2 demonstrate that perception quality and cognitive validity coexist in the proposed architecture. YOLOv12's high precision prevents spurious remediation, while DBN robustness ensures that legitimate misconceptions are detected early enough to guide adaptive branching. The synergy justifies treating computer-vision and learner-modelling pipelines as a single pedagogical instrument rather than isolated research artefacts, paving the way for data-centric, simulation-ready safety curricula.

Recommendations

To translate the prototype into a field-deployable teaching tool, future work should consolidate three mutually reinforcing vectors: data scalability, multimodal perception and cognitively informed adaptivity. First, synthetic-data pipelines ought to move beyond static Unity renders toward controllable diffusion models capable of fabricating rarely photographed hazards, such as live switch-gear arcing or unstable trench walls; preliminary studies show that such generators can raise small-class recall by up to six percentage points without inflating false alarms. Coupling these images with domain-randomized lighting schedules would further harden detectors against the night-shift and fog conditions documented as failure modes in mobile deployments. Second, decision-level fusion of video, inertial and BIM semantics has already cut occlusion-driven misses by roughly one-quarter while preserving real-time throughput; integrating such channels would allow the tutor to diagnose compound violations, for example height work without an anchor combined with tool over-reach, thereby aligning feedback with the complex causal reasoning required for PLO 2. Third, hierarchical DBN enriched with control nodes and hazard-severity weights consistently outperform flat tracing by 0.10–0.12 AUC, suggesting that next-generation models should embed forgetting curves and spaced-retrieval triggers to support longitudinal retention. Generative language–vision models already draft VR scenes that adapt complexity to mastery gaps, promising to cut instructional-design time and enable ubiquitous micro-learning at scale.

Conclusions

The study demonstrates that an anchor-free YOLOv12 detector, fine-tuned on a balanced hybrid corpus, and a six-node DBN calibrated on synthetic learner traces can jointly deliver millisecond-level adaptive feedback without head-mounted displays. Empirically, the detector attained 0.925 mAP and 56 FPS, outperforming YOLOv8 baselines by 3.5 points, while the DBN achieved 0.87 AUC and converged on mastery in fewer than ten interactions, benchmarks that mirror recent gains reported for synthetic-augmented vision and severity-weighted knowledge tracing. Educationally, the closed loop turns each hazard detection into Bloom-aligned evidence, replacing static quizzes with continuous embedded assessment and raising self-efficacy proxies by 35 % in simulated sessions,

consistent with immersive-VR studies on behavioral transfer. Limitations reside in the synthetic nature of learner traces and controlled camera viewpoints, yet prior DBN situational-awareness work suggests that well-parameterized networks transfer with minimal recalibration. By outlining a roadmap that scales synthetic data generation, multimodal fusion and hierarchical cognition, the research elevates AI-enabled safety tutoring from promising prototype to credible educational infrastructure. Even incremental gains in hazard recognition translate into tangible reductions in the 5 000-plus annual construction fatalities, furnishing a compelling moral and economic rationale for continued development.

References

Adil Raja, M., Loughran, R., & Mc Caffery, F. (2025). A review of performance of recent YOLO models on cholecystectomy tool detection. *Measurement: Digitalization*, *2–3*, 100007. https://doi.org/10.1016/j.meadig.2025.100007

Chen, H., Lei, Y., Xia, L., Deveci, M., Chen, Z.-S., & Liu, Y. (2025). Dynamic evaluation of the safety risk during shield construction near existing tunnels via a pair-copula bayesian network. *Applied Soft Computing*, *169*, 112583. https://doi.org/10.1016/j.asoc.2024.112583

Feng, F., Hu, Y., Li, W., & Yang, F. (2024). Improved YOLOv8 algorithms for small object detection in aerial imagery. *Journal of King Saud University - Computer and Information Sciences*, *36*(6), 102113. https://doi.org/10.1016/j.jksuci.2024.102113

Guo, X., Liu, Y., Tan, Y., Xia, Z., & Fu, H. (2024). Hazard identification performance comparison between virtual reality and traditional construction safety training modes for different learning style individuals. *Safety Science*, *180*, 106644. https://doi.org/10.1016/j.ssci.2024.106644

Huang, T., Hu, J., Yang, H., Hu, S., Geng, J., & Ou, X. (2025). Memory flow-controlled knowledge tracing with three stages. *Neural Networks*, *187*, 107292. https://doi.org/10.1016/j.neunet.2025.107292

Jiang, T., Fang, Y., Goh, J., & Hu, S. (2024). Impact of simulation fidelity on identifying swing-over hazards in virtual environments for novice crane operators. *Automation in Construction*, *165*, 105580. https://doi.org/10.1016/j.autcon.2024.105580

Loong, C. N., & Chang, C.-C. (2024). Control knowledge tracing: Modeling students' learning dynamics from a control-theory perspective. *Computers and Education: Artificial Intelligence*, *7*, 100292. https://doi.org/10.1016/j.caeai.2024.100292

Man, S. S., Wen, H., & So, B. C. L. (2024). Are virtual reality applications effective for construction safety training and education? A systematic review and meta-analysis. *Journal of Safety Research*, *88*, 230–243. https://doi.org/10.1016/j.jsr.2023.11.011

Seo, S., Park, H., & Koo, C. (2024). Impact of interactive learning elements on personal learning performance in immersive virtual reality for construction safety training. *Expert Systems with Applications*, *251*, 124099. https://doi.org/10.1016/j.eswa.2024.124099

Shabo, A., Guzdial, M., & Stasko, J. (1997). An apprenticeship-based multimedia courseware for computer graphics studies provided on the world wide web. *Computers & Education*, *29*(2), 103–116. https://doi.org/10.1016/S0360-1315(97)00031-6

Shi, Y., Du, J., & Zhu, Q. (2020). The impact of engineering information format on task performance: Gaze scanning pattern analysis. *Advanced Engineering Informatics*, *46*, 101167. https://doi.org/10.1016/j.aei.2020.101167

Xu, S., Sun, M., Fang, W., Chen, K., Luo, H., & Zou, P. X. W. (2023). A bayesian-based knowledge tracing model for improving safety training outcomes in construction: An adaptive learning framework. *Developments in the Built Environment*, *13*, 100111. https://doi.org/10.1016/j.dibe.2022.100111