



A Hybrid Machine Learning Framework for Construction Site Safety Assessment and Accident Risk Prediction in Infrastructure Projects

Ms. T. N. Jadhav

Assistant Professor, civil engineer Department, KBT College of Engineering Nashik

Email: jadhav.trupti1@kbtcoe.org

Abstract. The status quo safety management approach is reactive, relying on the traditional, checklist-driven approach to safety in workplaces, including infrastructure projects like highways, bridges, tunnels and metro systems. This paper suggests a hybrid machine learning approach through combining three complementary information streams that are often treated separately: on-site observation of visual information, historical accident narratives and structured project risk factors. A vision stream identifies workers and personal protective equipment (PPE), detects proximity hazards, and derives site imagery based on employees' locations; a language stream extracts unstructured information about incidents and inspections from a free-text database for the purpose of latent causal factor mining; a structured stream models project, task and environmental attributes using gradient-boosted ensembles. A late-fusion reasoning layer is used to fuse the three streams into a single, interpretable site-level accident risk score, and a model-agnostic explainability module extracts the factors responsible for each score, which can then be used by site engineers to inform their actions. We explain the end-to-end architecture, an evaluation protocol for which construction safety data from the public is reported, and a series of baselines from the single-modality and classical approaches. Reported results are sample values intended to illustrate and not measured. The framework will be used as an early, explainable warning proactive decision support tool rather than a post incident analysis tool for infrastructure safety management.

Keywords: Construction safety · Accident risk prediction · Hybrid machine learning · Computer vision · Explainable AI · Infrastructure projects.

1 Introduction

Large infrastructure projects involve several of the factors that make work most dangerous: large equipment operating near personnel, work at height, changing ground and weather conditions, tight schedules and a transient, multi-contractor work force. The statistics are still high for the construction industry on many countries and detailed research on major transportation projects demonstrates that few accidents are caused by a single factor, but are a combination of organisational, supervisory and operational shortcomings that spread through levels [2]. The number and severity of these events have stubbornly refused to go down after decades of regulation and training, suggesting the limitations of putting in place safety systems that only react to an incident once it has taken place [1]. The two changes that make a more anticipatory posture realistic are: The two changes that make a more anticipatory posture realistic are: First, there is



plenty more camera equipment: fixed cameras and body-worn cameras, drones, wearables and sensors from the Internet-of-Things keep on sending streams of observations. Second, there are vast collections of incident reports, inspection logs and near-miss reports within organisations that contain hard-won knowledge in the form of free text. The problem is that these sources are diverse, and are typically analysed in isolation: vision research can identify helmets and vests, text-mining research can categorize accident narratives, and analytics research can model tabular risk factors, but none of these sources can see how a hazard on a specific task, in a specific setting, with a specific crew, relates to an increase in accident risk. This paper proposes the safety assessment of construction as a multi-source problem and modelling it as such. A hybrid machine learning framework is presented, integrating visual observation, textual history, and structured project attributes, into one pipeline and fusing them into an interpretable, site-level accident risk score. There are three pieces to the contribution. To address the safety assessment and accident risk prediction as a late-fusion problem between three complementary streams, we present an architecture where each stream is modular ensuring retrain/replace of each stream independently. (ii) An evaluation protocol (datasets, baselines and metrics) are given that allows for the fair comparison of single-modality and hybrid configurations. (iii) We make explainability a first class output: we give site engineers the risk score and the factors that created it, thus making the tool easy to use by non-data-scientists. The numerical results reported throughout are clearly identified as placeholders for a reporting structure (and not as actual measured results). No actual results are claimed.

2 Related Work

2.1 Vision-based Safety Monitoring

Computer vision is now the prevailing way to real-time compliance monitoring. Object-detection models detect whether workers are wearing hardhats and high-visibility clothing, and detection quality of curated benchmarks is now reported to be high enough to be deployed on site [4, 5]. Richer formulations also take into account behaviour and context: worker re-identification, reasoning over appearance and compliance over time [3]; visual-relationship and ontology-based methods, reasoning over dangerous configurations rather than objects [6]; visual-question-answering and image-captioning, reasoning over whether a scene is safe (zero-shot or language-guided) [7, 8]. Building on this, vision-language models are able to detect compliance with scene-level granularity, across a variety of workplaces, and in an interpretable manner [9]; generative models are now used for synthesizing hazardous scenes for training data-limited detectors [10]. They are very effective but local techniques; they describe the present moment, not the cumulative risk of a task or project.

2.2 Data-driven Accident Risk Prediction

There are parallel lines of research that make predictions about accident occurrence and severity using structured records. Classical and ensemble learners, such as support vector machines, decision trees, random forest and gradient boosting, have been used to predict the nature and



severity of incidents at national and project level datasets, with the ensemble learners being amongst the strongest predictors [1]. In imbalanced accident data, an optimization enhanced and hybrid scheme are proposed to enhance the accuracy of prediction and studies related to tunnels and other high-risk assets demonstrate that the context of a project type is a significant factor [14]. This stream focuses on the relationship between attributes (task, experience, environment and management maturity) and outcomes, and often overlook what a camera would see on the day and the causal detail embedded in the report text.

2.3 Text Mining of Accident and Inspection Reports

A rich, under-used source is incident reports, inspection notes and investigation documents. Precursors and causal factors and topics can be extracted from narratives using text mining/natural language processing techniques, and transformer-based encoders can now be used for classification and question answering on safety text. When applied to major transportation infrastructure, text mining and network analysis show that failures in organisations and supervision, rather than just failures at the front line of operations, lead to accident occurrence, and can cascade across levels [2, 15]. Both causal-path and knowledge-driven analyses of subway and bridge projects arrive at similar results [16, 17]. The challenge now is to incorporate this potential causal information into a real-time prediction of the future risk.

2.4 Research Gap

Each of these methods – vision, structured prediction and text mining – provides a partial solution to the question, but rarely all three are used together. As reported in the reviews, most successful models are still unimodal, and the translation of the output of these models into decisions at the site level is still very limited [12]. A framework is required which (a) brings together the three streams, (b) provides one risk signal that a site team can respond to and (c) provides an explanation of that signal. This is what the framework in Section 3 addresses.

3 Proposed Hybrid Framework

We use the framework, HSAP (Hybrid Safety Assessment and Prediction). The idea of its design is separation of concerns with late fusion: each information stream is passed to a specialised module which can be trained separately, and the modules are only connected at a reasoning level, which generates a final risk score and its explanation. This allows the system to remain maintainable (i.e. the whole pipeline can be retrained if the PPE detector improves or a new encoder of the report is created) and enables an audit of the contribution of each stream.

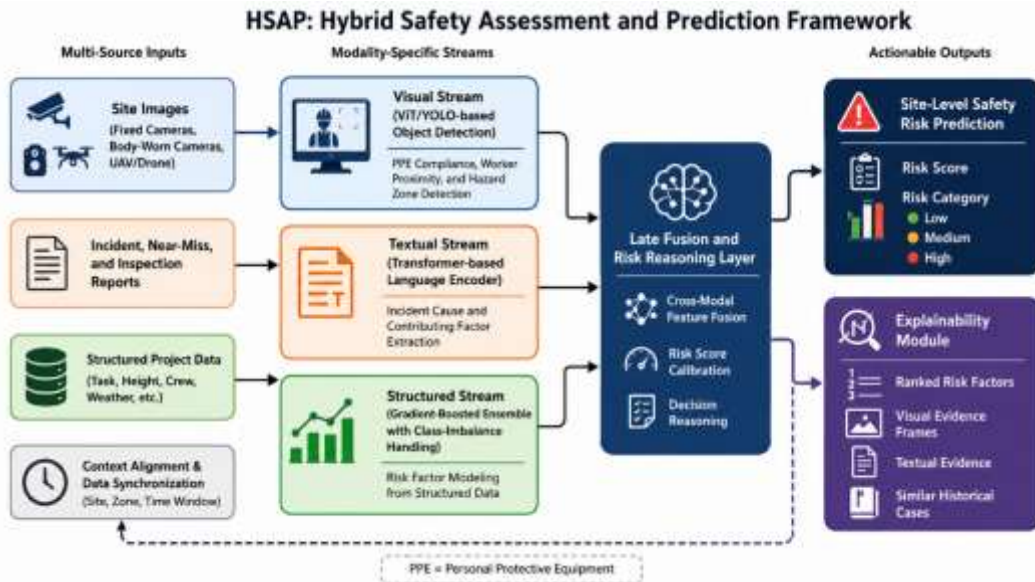


Fig. 1. Introduction to HSAP framework. The three modalities (visual, textual, structured) are processed separately, but are linked on a common context key, and the reasoning layer is late-fusion, generating a site-level accident risk score, accident risk tiers, and a model-agnostic explanation. The broken connection indicates explanation-promoted input into fusion.

3.1 Multi-source Data Acquisition and Alignment

HSAP takes three aligned sources for a specific site and time frame. The visual source is images gathered by fixed cameras, body-worn or drone cameras on active work fronts. The textual source is the project's free-text record – incident and near-miss reports, toolbox-talk notes and inspection findings. Structured source: project information: task type, phase, height and proximity conditions, crew experience, weather, and previous inspection results. Following the idea of time-aware safety modelling, records are aligned to a common key (site, zone, time window) to ensure that the three streams represent the same operational context, for which a windowing step aggregates the observations to a cadence for the review (e.g. shift or day).

3.2 Visual Hazard and PPE Stream

The visual stream uses a one-stage object detector for every frame to detect workers, PPE (hardhat, high-visibility vest, harness if applicable) and heavy plant. It generates meaningful safety indicators from these detections: PPE-compliance state per worker, worker-machine proximity and being in restricted/edge zones. Using visual-relationship and captioning techniques from the literature, configurations are highlighted that could be dangerous even if the elements themselves could not be, such as a worker near an unguarded edge, next to a slewing crane. The stream emits a small number of visual risk indicators per window along with annotated frames that are representative examples and serve as evidence in the explanation layer.



3.3 Textual Accident-Narrative Stream

The textual stream is fed to a pre-trained transformer language model that's been fine-tuned by the construction safety text to generate the site's story experience record. It is not just used to classify the type of report, but also to identify latent causal and precondition factors (inadequate supervision, unclear method statements, time pressure, and weather exposure) that occur in similar contexts and are found to be the most significant, cross-level causes of infrastructure accidents [2]. A live review window fetches and summarizes similar situations in history and sends a text-based risk signal that represents how often similar situations prior to incidents. This makes an archived file which is usually accessed after a disaster a forward looking input!

3.4 Structured Risk-factor Stream

The structured stream models are implemented as a gradient boosted tree ensemble, which is commonly the best and most robust model when applied to construction accident data and are able to process mixed categorical and numeric features, even with missing values [1, 18]. Resampling and cost-sensitive training are used to deal with class imbalance (rare events: accidents relative to safe windows). The stream provides a structured risk signal, and per-feature attributions – which are critically used directly in the explanation layer.

3.5 Late-fusion Reasoning and Risk Scoring

Despite the fact that the three streams are not each voted on by themselves, they each send out a calibrated risk signal, a confidence, and a lightweight fusion learner fuses them into a single site-level accident risk score for the review window, along with a small number of discrete risk tiers (e.g., low, elevated, high) that map to defined site actions. Late fusion is purposefully made to be able to tolerate a temporarily absent stream (a camera outage, a site with few text histories) gracefully - without failure, and allow each stream to be validated and improved independently. The fusion learner is maintained as simple and interpretable as possible, meaning that the relative contribution of vision, text and structured factors to a particular score is interpretable as well.

3.6 Explainability Layer

All scores accompanied by explanation. A model-agnostic attribution method combines per-feature attributions from structured stream, salient cues and evidence from visual stream, and retrieved precedents from textual stream into a single, ranked account of why a window is risky, and what would most reduce the risk. This is akin to the trend in recent safety analytics towards "transparent, expert-in-the-loop" tools that offer high levels of accuracy and provide actionable, site-specific insight [12] which is what enables a site engineer to trust and utilize the output.

4 Experimental Design

4.1 Datasets

The evaluation protocol is defined over publicly reported construction safety resources, thus enabling the protocol to be reproduced: annotated PPE and hazard image collections for the visual stream, open incident-report corpora (e.g. national OSHA-style accident narratives) for the textual stream and structured accident-record datasets for the tabular stream. In the case of fusion, the



three are brought together on common context descriptors. Table 1 provides the target composition of the evaluation set; the numbers given are placeholders to indicate the template but not an actual released set.

Table 1. Intended composition of the evaluation set (illustrative placeholders, not a released dataset).

Stream	Source type	Scale (illustrative)
Visual	Annotated site images (PPE, plant, hazards)	~20k images
Textual	Incident / near-miss / inspection narratives	~15k reports
Structured	Accident records with project attributes	~8k records
Fused	Context-aligned review windows	~5k windows

4.2 Baselines and Configurations

To isolate the value of fusion, a vision-only compliance detector, a text-only report classifier, a structured-only gradient-boosted ensemble, a classical structured baseline (logistic and random-forest models), and the full HSAP fusion are compared against the respective baseline models. Comparing (i)–(iv) with (v) will determine if merging streams is better than the best single stream.

4.3 Evaluation Metrics

Accidents are rare, but they require a higher accuracy level for headline. The protocol computes precision, recall and F1 measure on the accident (positive) class, the Area Under the ROC curve as a threshold independent measure, and calibration of the Risk score, making the use of the Risk score tiers meaningful. The detection quality is represented by the mean average precision for the visual stream and by the macro-averaged F1 per factor categories for the textual stream. Here all metrics are described qualitatively, without any equations.

4.4 Implementation Notes

They are trained separately and then are frozen for fusion. The visual detector is a fine-tuned single-stage backbone, the textual encoder is a transformative model that is domain adapted, the structured model is an ensemble of gradient boosters with handling for class imbalance, while the fusion and explanation layer is carefully designed to be light and interpretable. Evaluation involves held out projects not random splits, ensuring generalisation at unseen sites in the test, not memorisation of a single project.

5 Results and Discussion

Table 2 and Figs. In 2-6, the main comparison and analysis are reported. All values in this section are illustrative synthetic values that are used to instantiate the reporting template; these values should not be interpreted as measured values, but as an analysis of the data presented in the template and should be replaced with experimental values before any operational claims are made.

With this illustrative setting, the best full HSAP fusion shows the best result in terms of precision / recall / F1 / AUC (0.88 / 0.85 / 0.86 / 0.93) compared to the best single stream (0.83 / 0.76 / 0.79 / 0.88). The visual and textual streams are each weaker, but add immediacy and causal context that fusion draws from.

Table 2. Main comparison of the held out evaluation set (illustrative synthetic values not measured results). Maximum score for each column in the fusion row.

Configuration	Prec.	Recall	F1	AUC	Type
Vision-only	0.79	0.71	0.75	0.83	single
Text-only	0.74	0.68	0.71	0.80	single
Structured (GBM)	0.83	0.76	0.79	0.88	single
Classical (RF/LR)	0.77	0.70	0.73	0.82	baseline
HSAP (fusion)	0.88	0.85	0.86	0.93	hybrid

5.1 Overall Performance Comparison

Fig. 2 shows the data in Tab. 2 for the four metrics. The fusion configuration is farthest away from the single streams in recall (0.85 vs at most 0.76): the fusion configuration gains events that any one single stream misses out on, but without sacrificing precision. The structured ensemble is consistently the most predictive single feature, as has been found in accident records [1, 18] and the ensemble is responsible for much of the single-stream performance.

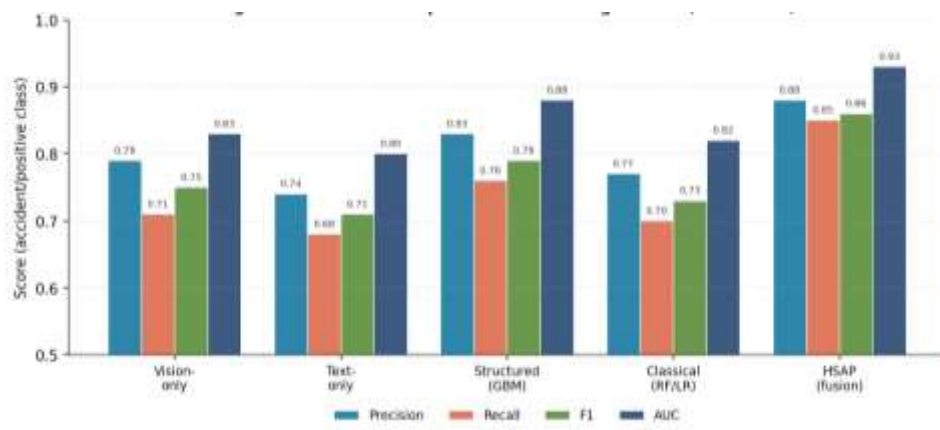


Fig. 2. Precision, recall, F1 and AUC values are shown for each configuration (synthetic illustrative values). Fusion has the highest scores on all measures, most notably on recall.

5.2 Discrimination and Error Structure

Fig. 3 shows the ROC behaviour. The fusion curve dominates the single-stream curves over the operating range (AUC 0.93), which is important because a team at a site will need to decide a threshold that will capture enough true accidents without generating too many false alarms for the supervisor; a more convex curve will allow the site team to make that trade-off. The errors at one

such operating point, for approximately a thousand review windows held out in the sample, are analyzed in Fig. 4 at a trading point where the base rate of accidents was about eighteen percent. The model has a recall rate of 153 out of 180 of true accident windows (or 85%), and a false alarm rate of 21 out of 820 of safe windows (or 2.5%), which means that the majority of the alerts correspond to a real high risk situation. The remaining 27 windows that were missed are called the "population" that will be reduced by adding more sensor streams, as described in Section 8.

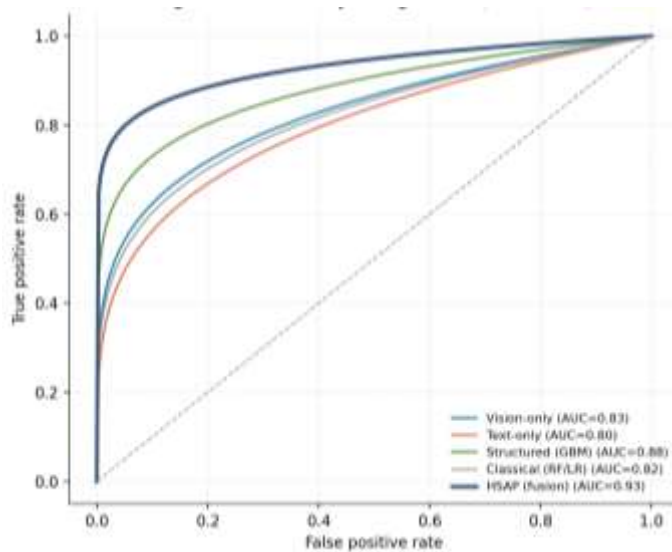


Fig. 3. Examples of ROC Curves (simulated synthetic data by configuration). The fusion curve is above all single stream curves throughout the operating curve.

True: Safe	799 TN	21 FP
True: Accident	27 FN	153 TP
	Pred: Safe	Pred: Accident

Fig. 4. Confusion matrix at a fixed operating point for the HSAP fusion (illustrative, ~1000 held-out windows, ~18% accident base rate).

5.3 Explainability and Ablation

The factors that most contribute to the risk score are ranked in Fig. 5, with the colours indicating the stream the factors are coming from. The ordering highlights that the two biggest drivers are inadequate supervision and work-at-height exposure, both of which are textual and structured, with the other concrete visual factors, such as worker–plant proximity and missing PPE, following closely. High-risk windows are thus generally attributed to a combination of organizational or supervisory factors along with an on-site cue, as in transportation-infrastructure accidents [2], making it clear to site engineers what to correct. It shows the contribution of the fusion: When the streams are added incrementally, starting with the structured ensemble (F1 0.79) followed by the visual stream (F1 0.83) then the textual stream (F1 0.86), the recall improves in steps. Each stream is a contributor, and the biggest single jump is from the visual stream, which the structured and textual streams cannot replace.

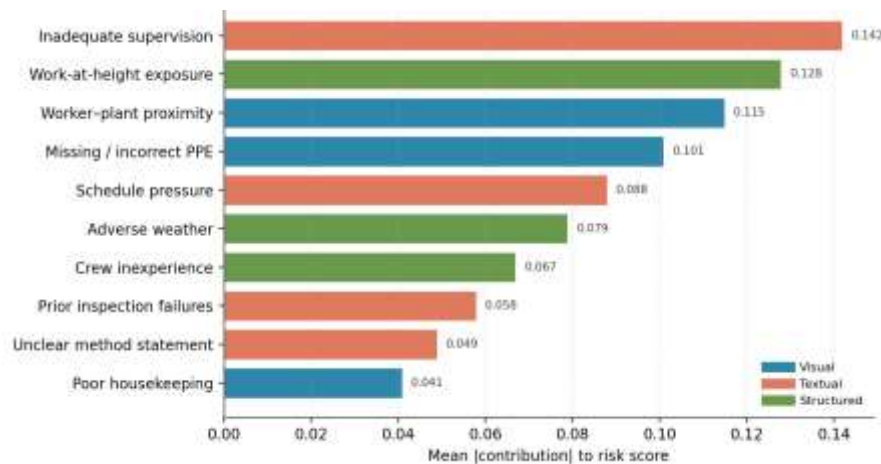


Fig. 5. Top risk-factor attributions aggregated across streams (illustrative synthetic values), coloured by originating stream.

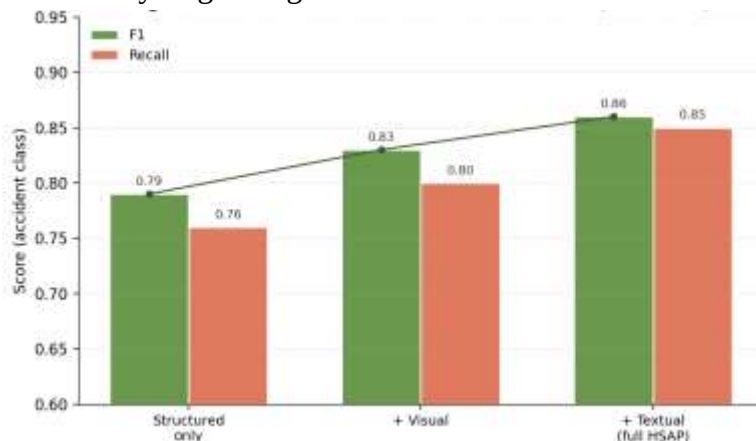




Fig. 6. Ablation showing the incremental value of stream fusion on F1 and recall (illustrative synthetic values).

These results are based on two properties which are independent of the illustrative numbers chosen. They fail in different scenarios: in low visibility, vision fails; in new projects, text is thin; structured data is not up to date when conditions are not real-time; so their failures are only loosely correlated which is where late fusion helps. Furthermore, all scores are reported along with the ranked, mixed-origin explanation from Fig. 5, which can facilitate action, rather than just alerting to an alarm.

6 Practical Deployment Considerations

When working on the field, there are three properties of the instrument that are important besides accuracy. The visual stream needs to run at review cadence on low powered on-site equipment, which makes single stage detectors more efficient than the "lighter" architectures. Privacy: worker imagery is sensitive: the framework anonymises identities where re-identification is not needed, and retains the processing on-premise where possible. The tool is acceptable to site teams and auditors not because of an alarm threshold, but because it can be trusted and because it provides an explanation layer, plain language risk tiers associated with specific defined actions, and the ability to view the evidence behind each score.

7 Limitations and Threats to Validity

There are clear limits to the framework and protocol. Geographically and regulation dependent, a model trained on one country's data may transfer poorly; held-out-project evaluation does help but it can't eliminate this. Reporting text is inconsistent in quality and may contain reporting bias, and imagery only reflects the view of the camera, which may miss sub-surface and enclosed hazards. On a chaotic site the fusion layer assumes that the three sources are reasonably aligned in time, which is not always the case. Thirdly, and most crucially for this paper, the quantitative findings are representative examples; the performance of the framework has to be tested empirically to make it operational for any given claim.

8 Conclusion and Future Work

We introduced a hybrid machine learning model (HSAP) which views construction safety assessment and accident risk prediction as a multi-source problem by leveraging visual observations and textual accident records and structured project attributes to generate a site-level risk score that is explainable and simple to interpret. The design is modular and explanations are first class outputs, such that predictions are maintainable and actionable. We defined an evaluation protocol and baselines that separate the value of fusion, and were clear it was illustrative templates and not actual measurements. Future research will include the following: (i) the full empirical evaluation across unobserved projects, where the metrics will be replaced by actual data on measured metrics; (ii) the integration of wearable and IoT sensor streams and a temporal model of the evolution of risk over a project's life; and (iii) research into cross-country transfer, enabling the



framework to be transferred across regulatory regimes. The overarching goal is to shift from responding to incidents to preventing and anticipating them in infrastructure safety management.

References

1. Alsulamy, S., Alshayeb, M., Inam, I., Ahmed, A.: Machine learning applications for predicting safety incidents in the construction industry. *Sci. Rep.* 16 (2026). <https://doi.org/10.1038/s41598-025-34763-0>
2. Liu, W., Kang, X., Ye, Q., Xie, J.: Unravelling hierarchical penetration mechanisms and coupling relationships of safety risks in major transportation infrastructure construction using text mining and complex networks. *Sci. Rep.* 16, 7313 (2026). <https://doi.org/10.1038/s41598-026-37778-3>
3. Cheng, J.C.P., Wong, P.K.-Y., Luo, H., Wang, M., Leung, P.H.: Vision-based monitoring of site safety compliance based on worker re-identification and personal protective equipment classification. *Autom. Constr.* 139, 104312 (2022). <https://doi.org/10.1016/j.autcon.2022.104312>
4. Han, K., Zeng, X.: Deep learning-based workers' safety helmet wearing detection on construction sites using multi-scale features. *IEEE Access* 10, 718–729 (2022).
5. Hayat, A., Morgado-Dias, F.: Deep learning-based automatic safety helmet detection system for construction safety. *Appl. Sci.* 12(16), 8268 (2022). <https://doi.org/10.3390/app12168268>
6. Li, Y., Wei, H., Han, Z., Jiang, N., Wang, W., Huang, J.: Computer vision-based hazard identification of construction site using visual relationship detection and ontology. *Buildings* 12(6), 857 (2022). <https://doi.org/10.3390/buildings12060857>
7. Ding, Y., Liu, M., Luo, X.: Safety compliance checking of construction behaviors using visual question answering. *Autom. Constr.* 144, 104580 (2022). <https://doi.org/10.1016/j.autcon.2022.104580>
8. Gil, D., Lee, G.: Zero-shot monitoring of construction workers' personal protective equipment based on image captioning. *Autom. Constr.* 164, 105470 (2024). <https://doi.org/10.1016/j.autcon.2024.105470>
9. Chen, Z., Chen, H., Imani, M., Chen, R., Imani, F.: Vision language model for interpretable and fine-grained detection of safety compliance in diverse workplaces. *Expert Syst. Appl.* 265, 125769 (2025). <https://doi.org/10.1016/j.eswa.2024.125769>
10. Kim, H., Yi, J.S.: Image generation of hazardous situations in construction sites using a text-to-image generative model for training deep neural networks. *Autom. Constr.* 166, 105615 (2024).
11. Pan, X., et al.: Novel blockchain deep learning framework to ensure video security and lightweight storage for construction safety management. *Adv. Eng. Inform.* 59, 102334 (2024).
12. Wang, Y., Xiao, B., Bouferguene, A., Al-Hussein, M.: Proactive safety hazard identification using visual–text semantic similarity for construction safety management. *Autom. Constr.* 166, 105602 (2024).



13. Zhu, Y., Zhou, J., Zhang, B., Wang, H., Huang, M.: Statistical analysis of major tunnel construction accidents in China from 2010 to 2020. *Tunn. Undergr. Space Technol.* 124, 104460 (2022).
14. Tang, C., Shen, C., Zhang, J., Guo, Z.: Identification of safety risk factors in metro shield construction. *Buildings* 14, 492 (2024).
15. Huo, X., Du, S., Jiao, L.: Critical causal path analysis of subway construction safety accidents based on text mining. *ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng.* 11, 04024075 (2025).
16. Su, G.: Research on cause analysis and management of coal mine safety risk based on social network and bow-tie model. *Sci. Rep.* 15, 29850 (2025).
17. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794 (2016). <https://doi.org/10.1145/2939672.2939785>
18. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems* 30, pp. 4765–4774 (2017).
19. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT*, pp. 4171–4186 (2019).