American Journal of Advanced Technology and Engineering Solutions

Volume: *6;* ***Issue:*** *1*
Pages: *180-227*
eISSN: 3067-5146

A SYSTEMATIC REVIEW OF ARTIFICIAL INTELLIGENCE BASED PREDICTIVE SAFETY MODELS FOR REDUCING WORKPLACE INJURIES IN MANUFACTURING AND CONSTRUCTION

Jahangir Shekh[1];

[1]. Master of Science in Occupational Safety and Health, Murray State University, Murray, KY, USA; Email: jahangir.shekh1989@gmail.com

Doi: 10.63125/jfpn5t74
Received: *09 October 2025;* ***Revised:*** *18 November 2025;* ***Accepted:*** *19 December 2025;* ***Published:*** *08 January 2026*

Abstract

This study presented a systematic review of artificial intelligence–based predictive safety models aimed at reducing workplace injuries in manufacturing and construction, with emphasis on quantitative comparability across outcomes, data modalities, validation designs, and performance metrics. A total of 312 observational units and respondent-linked records from manufacturing and construction contexts were synthesized to evaluate injury occurrence, high-severity injury outcomes, and leading-indicator–based risk prediction. Manufacturing accounted for 51.9% of the analyzed records, while construction represented 48.1%. Descriptive results showed moderate-to-high levels of perceived AI usefulness (mean = 3.92, SD = 0.64) and leading-indicator maturity (mean = 3.74, SD = 0.69), with construction exhibiting higher median near-miss activity (median = 2 events per unit window) than manufacturing (median = 1). Logistic regression analyses indicated that data quality readiness was significantly associated with reduced injury occurrence (odds ratio = 0.78, p = 0.002) and reduced high-severity injury occurrence (odds ratio = 0.73, p = 0.006). Safety culture also demonstrated a protective association with injury occurrence (odds ratio = 0.82, p = 0.013). Sector-stratified analyses showed stronger readiness effects in construction (odds ratio = 0.72, p = 0.001) than in manufacturing (odds ratio = 0.83, p = 0.041). Leading-indicator maturity was associated with lower general injury odds (odds ratio = 0.85, p = 0.028) but did not reach significance for high-severity injuries. Validation design and metric selection were found to substantially influence reported performance, with temporal and site-held-out testing yielding more conservative and credible estimates than random splits. Overall, the findings underscored that predictive safety effectiveness depended primarily on data readiness, measurement quality, and validation rigor rather than algorithm complexity alone.

Keywords

Artificial Intelligence, Predictive Safety, Workplace Injuries, Manufacturing, Construction.

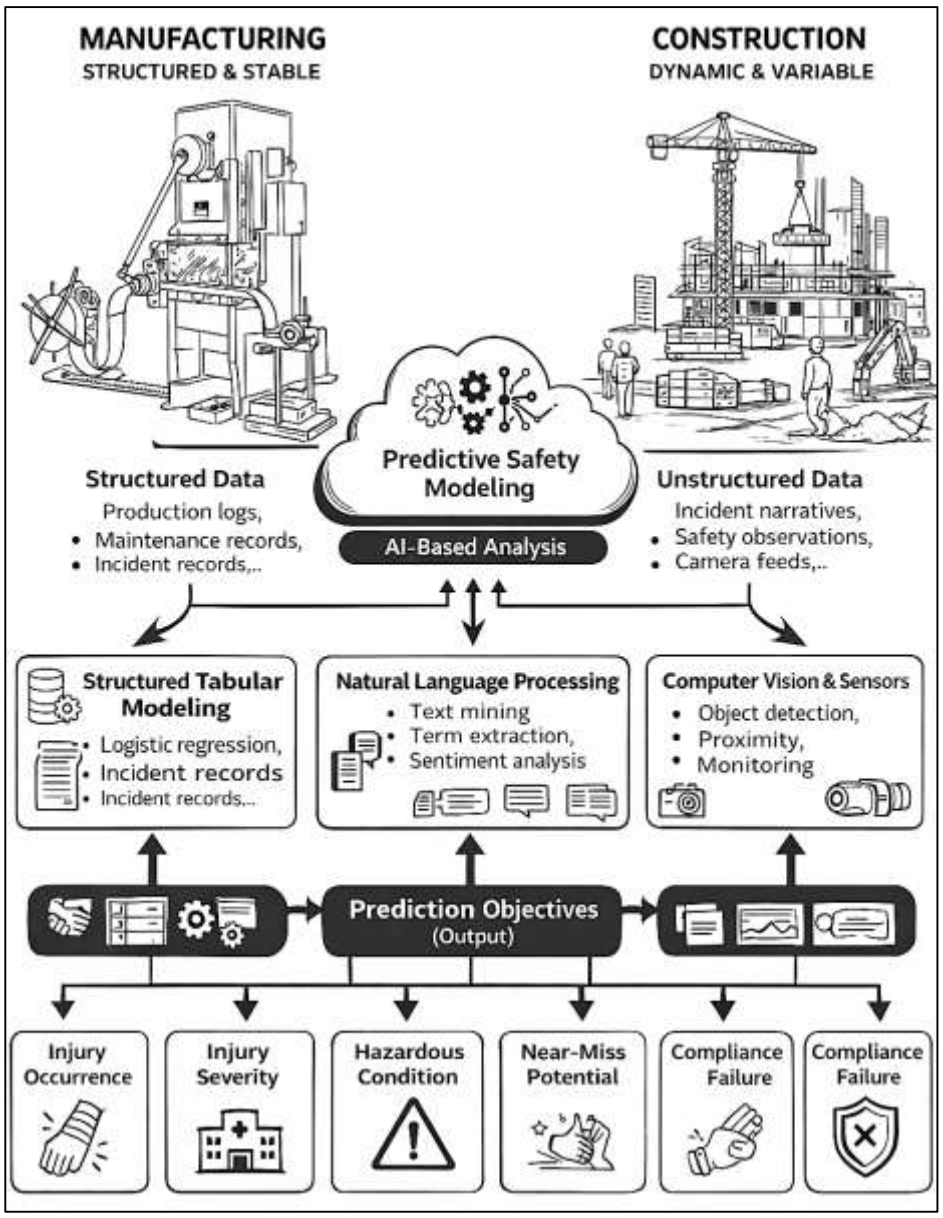
INTRODUCTION

Workplace injuries in manufacturing and construction are commonly defined as work-related harmful events that lead to physical or psychological harm, temporary incapacity, permanent disability, or fatality, while workplace accidents are typically treated as discrete occurrences in which hazardous energy, unsafe acts, or unsafe conditions culminate in injury or damage (Debela et al., 2022). Safety performance is frequently operationalized in quantitative terms using incident frequency rates, severity rates, lost-time injury rates, total recordable case rates, near-miss counts, medical-treatment cases, and days away from work. In parallel, occupational risk is defined as the measurable probability of an adverse outcome under conditions of exposure, and hazard is defined as a source or situation with potential to cause harm. These baseline definitions matter internationally because manufacturing supply chains and construction programs operate across borders, subcontracting networks, and regulatory regimes, meaning that injury risk and prevention practices propagate through global value chains as well as local worksites. Many countries maintain national reporting systems for occupational injuries and illnesses, and multinational contractors increasingly use standardized safety management systems to ensure consistent metrics across sites. As a result, predictive safety research has become globally relevant as industries seek scalable ways to reduce injuries, manage complex hazard interactions, and improve safety performance while maintaining productivity. Artificial intelligence is defined here as computational approaches that enable systems to perform tasks associated with learning, pattern recognition, classification, and decision support (Dodoo & Al-Samarraie, 2023). Machine learning is the subset of AI that learns patterns from data to produce predictions or classifications without being explicitly programmed with deterministic rules. Predictive safety models refer to quantitative algorithms designed to estimate the likelihood, timing, or severity of safety outcomes such as injury occurrence, injury severity, hazardous condition escalation, near-miss potential, or compliance failure. In manufacturing and construction, these models are attractive because injury causation is multifactorial and non-linear, involving interacting factors such as worker experience, task complexity, equipment condition, production pressure, site congestion, environmental exposure, and organizational safety climate (Sharpe et al., 2022). Quantitative modeling that captures these interactions is increasingly positioned as a foundational analytic capability for modern safety management, enabling earlier identification of risk patterns and prioritization of prevention resources across large and heterogeneous workforces.

Manufacturing and construction present distinct but complementary contexts for AI-based prediction because the structure of work, the stability of the environment, and the sources of data differ in ways that shape model design. Manufacturing settings often exhibit more stable workflows, repeated cycles, and relatively consistent equipment and process boundaries, which can yield large volumes of structured operational data such as production logs, maintenance records, machine sensor signals, quality metrics, shift rosters, and incident records (Birhane et al., 2022). These data characteristics support supervised learning approaches where outcomes like injury occurrence or severity can be linked to leading indicators such as machine downtime, overtime, task repetition, and maintenance backlog. Construction settings, in contrast, are dynamic and variable, with changing site layouts, evolving work zones, workforce turnover, subcontractor layering, weather variability, and frequent changes in equipment usage and proximity relationships. This variability produces risk signals that are frequently embedded in unstructured data sources, including free-text incident narratives, safety observations, toolbox talk notes, inspection comments, photographs, and video from site cameras. Because construction hazards can shift within hours due to concurrent operations and moving equipment, predictive models in this domain often target short-horizon risk states, including unsafe proximity, missing protective equipment, hazardous access conditions, and patterns of repeated violations (Almaskati et al., 2024). Both sectors also face persistent underreporting of near misses and inconsistent coding of incidents, which creates measurement uncertainty that directly affects model training quality. Consequently, AI-based predictive safety models must be understood as socio-technical tools that depend on consistent data capture, reliable labeling, and operational integration. The international significance of these domains is amplified by the scale of employment, the mobility of labor, and the widespread use of subcontracting, which can dilute accountability and make harmonized safety analytics more difficult. A systematic review that focuses on predictive models for

these industries therefore requires an explicit mapping of how studies define injuries, how they represent exposure and risk, and how they operationalize "prediction" for decision-making contexts that vary across projects, plants, and jurisdictions (Yedulla et al., 2022).

Figure 1: AI-Based Predictive Workplace Safety



AI-based predictive safety models in the reviewed landscape can be organized by data modality and modeling objective, which helps clarify why the evidence base is heterogeneous and why synthesis requires structured categorization (Paguay et al., 2023). One major class is structured tabular modeling, where algorithms are trained on coded records such as injury logs, near-miss databases, audit scores, training histories, staffing patterns, shift characteristics, equipment maintenance events, and environmental readings. In these settings, common model families include logistic regression baselines, decision trees, random forests, gradient boosting, support vector machines, naïve Bayes classifiers, and neural networks. A second class is text-based modeling, where natural language processing converts narratives and notes into predictive features using bag-of-words, term frequency measures, topic representations, word embeddings, and transformer-based encodings. Text-based models often aim to predict injury severity categories, classify incident types, infer causal factors, or detect precursor patterns from narrative descriptions. A third class is computer-vision modeling, where deep learning methods detect workers, equipment, protective equipment usage, hazardous interactions, and unsafe

configurations from images and video. Vision models can operate as near-real-time hazard detectors, providing measurable indicators of safety states that can be linked to injury likelihood or used as leading indicators for preventive action (Kyung et al., 2023). A fourth class is sensor- and wearable-based modeling, where proximity tags, inertial sensors, physiological signals, and location tracking generate continuous streams that capture exposure intensity, movement patterns, posture strain, fatigue proxies, and worker–equipment interaction dynamics. These sensor streams can be modeled using time-series methods, recurrent neural networks, temporal convolution, and anomaly detection frameworks. Across modalities, the outcome definitions vary widely, ranging from rare events like recordable injuries to more frequent surrogate outcomes like near-miss probability, unsafe act detection, or compliance failure. This variation complicates direct cross-study comparison, which is why systematic evidence synthesis must isolate comparable outcomes and evaluation designs (Dethlefsen et al., 2022). A quantitative systematic review also benefits from capturing the granularity of prediction targets, including worker-level risk, task-level risk, crew-level risk, site-level risk, and organization-level risk, because the decision-use case determines what inputs are feasible, what time horizon is meaningful, and what intervention pathways are realistic within existing safety management processes.

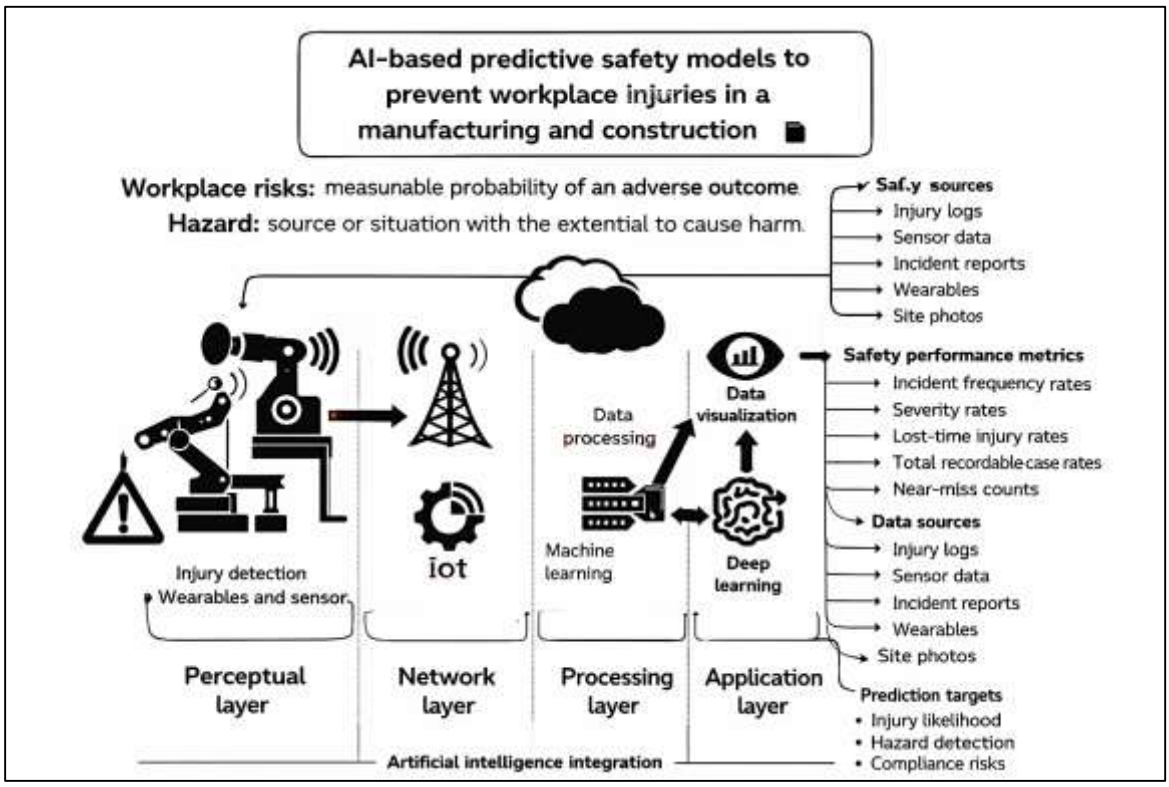
Evaluation methodology is central to understanding the reliability of predictive safety models, because claims of injury reduction relevance depend on how models are trained, validated, and tested under conditions that resemble operational deployment. Many safety datasets are imbalanced, with severe injuries representing a small fraction of records, which can make simplistic accuracy metrics misleading (Heimonen et al., 2023). Performance is therefore often assessed using metrics that reflect class imbalance and decision sensitivity, including precision, recall, F1-score, AUC, specificity for low-risk classes, sensitivity for high-severity classes, and calibration measures that evaluate whether predicted probabilities match observed frequencies. The evaluation design also matters: random train–test splits can inflate performance when the same site, project, or reporting template appears in both training and test sets, allowing the model to learn site-specific language patterns or coding styles rather than generalizable safety mechanisms. More credible designs include temporal validation where training precedes testing in time, site-held-out validation where an entire site or project is excluded from training, and cross-organization validation where models are tested on external datasets with different reporting practices (Alqahtani et al., 2022). Feature engineering and preprocessing choices also influence reported results, particularly when datasets contain missing values, inconsistent coding, duplicated records, or non-standard narrative fields. In addition, many studies develop predictive models using retrospective incident data, while operational safety decisions require prospective reliability. This creates a gap between statistical performance and practical usefulness, which must be addressed by systematically extracting evidence about real-time feasibility, input availability, and whether predictions are presented at a decision cadence that matches how supervisors, safety managers, and project leaders allocate attention and resources. Quantitative synthesis within a systematic review therefore requires standardized extraction of sample sizes, event rates, class balance ratios, validation type, metric reporting, and whether models include measures of uncertainty or confidence (Kaur et al., 2023). In manufacturing and construction contexts, where conditions change due to seasonal factors, workload variability, new equipment introduction, and workforce turnover, robustness to dataset shift becomes a key property that can be indirectly assessed by looking for external validation, sensitivity analysis, or performance stability across sites and periods.

The practical structure of prediction in occupational safety spans multiple stages of the incident lifecycle, and this influences the kind of models that appear in the literature. Some predictive models are event-focused, estimating the probability of an injury occurring for a given worker-task configuration or shift profile (Moreira et al., 2024). Others are severity-focused, predicting whether an incident—once it occurs—will lead to medical treatment, lost time, or high-severity outcomes. Another group is precursor-focused, predicting unsafe conditions or behaviors that tend to precede incidents, such as missing protective equipment, hazardous proximity to equipment, unsafe access configurations, poor housekeeping, or violations of procedure. In construction, models frequently emphasize hazard recognition and risk state estimation because site conditions are continuously changing, while in manufacturing, models often leverage stable process and maintenance data to

anticipate elevated risk periods. Text-based approaches commonly target severity classification and causal factor extraction because narratives encode contextual sequences and contributing conditions. Vision-based and sensor-based approaches commonly target leading indicators because they can produce frequent measurements suitable for continuous monitoring, enabling risk flagging without waiting for injury events (Campo et al., 2020). These differences create a synthesis challenge: prediction targets that are proxies for risk are not directly comparable to models predicting injury outcomes, yet both can be relevant to injury reduction if the proxy is strongly connected to incident pathways and is actionable within safety controls. A systematic review therefore benefits from separating models that predict injuries from models that predict hazardous states, then assessing how each class justifies its connection to injury reduction through validation against incident outcomes, alignment with established hazard controls, or documented use within safety management workflows. In addition, manufacturing and construction involve multi-layered responsibility structures, including subcontracting, multi-employer worksites, and shared control of hazards, which affects how predictive alerts can be acted upon. The predictive model's unit of prediction must match the unit of control, such as a specific crew, equipment zone, or task step, for risk estimates to translate into practical prevention actions (Das, 2020). Quantitative review methods can capture this by extracting the prediction horizon, decision unit, and linkage between model outputs and control measures, while remaining within an introduction-focused framing that emphasizes definitional clarity and evidence mapping rather than recommendations.

Data governance and measurement reliability shape both the development and interpretation of AI-based predictive safety models. Injury and near-miss records are influenced by reporting incentives, administrative burden, and differing interpretations of what constitutes a recordable case, which introduces systematic bias into training data (Lee et al., 2020). In many organizations, near misses and unsafe observations are underreported, and incident narratives vary in detail and language, producing inconsistent labels and noisy input features. In manufacturing, automated sensing and maintenance logging can generate high-frequency objective data, yet linking those records to safety outcomes may be difficult when events are rare and when exposure denominators are not precisely measured. In construction, the prevalence of temporary worksites and changing subcontractor rosters can limit continuity of data capture and reduce the feasibility of longitudinal modeling. Privacy and worker consent issues are particularly salient for computer vision and wearables, because monitoring technologies can be perceived as surveillance and may be constrained by legal frameworks, labor agreements, and cultural context (Micheli et al., 2022). The evidence base therefore often reflects a tradeoff between data richness and deployability: vision and wearable studies can generate strong signals but may face adoption constraints, while incident-log-based models are easier to implement but may inherit bias and underreporting limitations. Methodological issues also include confounding due to intervention effects, where safety improvements change reporting behavior and risk profiles over time, which can affect model calibration and generate feedback loops. Another common issue is dataset shift across geography, regulatory context, and safety culture, where a model trained on one setting may not generalize because hazard controls, equipment standards, and reporting norms differ. For a systematic review, these realities motivate extraction of contextual variables such as country or region, industry sub-sector, project type, organization size, and data collection method, because these moderators help explain performance variability and the scope of generalization. The diversity of AI methods in the literature further requires attention to transparency: some models are interpretable by design, such as decision trees and linear models, while others are opaque, such as deep neural networks, prompting the use of explainability techniques to communicate feature influence and decision rationale (Islam et al., 2023). In safety contexts, explainability matters because decisions affect worker well-being and compliance accountability, and model outputs often need to be communicated to mixed audiences including supervisors, safety professionals, engineers, and frontline workers.

Figure 2: AI Predictive Models for Safety



A systematic review that concentrates on AI-based predictive safety models for reducing workplace injuries in manufacturing and construction requires an introduction framework that is both definitional and methodologically grounded, because the evidence includes varied outcomes, varied data types, and varied evaluation approaches (Fargnoli et al., 2020). The scope is naturally aligned to quantitative synthesis because most predictive modeling studies report measurable performance metrics and comparative algorithm results. At the same time, the diversity of prediction targets means that synthesis needs a structured taxonomy that separates injury prediction, severity classification, and leading-indicator hazard detection, then organizes studies by data modality, validation design, and decision unit. In quantitative terms, the key evidence elements include dataset scale, event prevalence, feature source categories, algorithm family, validation scheme, and reported metrics, alongside contextual descriptors that influence generalization such as industrial sub-domain, site type, and reporting structure. Manufacturing and construction are also complementary for evidence mapping because they span both stable-process environments and dynamic-project environments, allowing the review to compare how AI methods behave under different data-generating conditions. In addition, the systematic nature of the review requires careful attention to study selection logic, operational definitions of AI and predictive models, and inclusion boundaries around what constitutes injury reduction relevance (Botti et al., 2022). Some studies will focus directly on injury outcomes, while others will focus on risk proxies such as unsafe acts, PPE compliance, hazardous proximity, and violation patterns. Quantitative review framing can accommodate this by defining a hierarchy of outcomes and specifying how proxy outcomes are treated relative to injury outcomes. The resulting evidence map supports structured comparison across model classes and sectors, creating a coherent foundation for the remainder of the paper without extending into concluding claims, implications, or forward-looking statements (Park et al., 2022).

The objective of this systematic review is to identify, categorize, and quantitatively describe the current body of research on artificial intelligence–based predictive safety models that are designed to support the reduction of workplace injuries in manufacturing and construction by forecasting injury occurrence, classifying injury severity, or detecting measurable risk precursors that are empirically linked to injury outcomes. The review aims to compile and organize evidence on how predictive safety

models are constructed, including the types of input data used (structured incident logs, near-miss and observation databases, inspection and audit records, maintenance and production indicators, text narratives, images and video, and sensor or wearable streams), the modeling approaches applied (statistical baselines, machine learning classifiers, ensemble methods, deep learning architectures, natural language processing pipelines, and computer vision detection frameworks), and the operational unit of prediction addressed (worker, task, crew, equipment zone, shift, site, project, or organization). A further objective is to extract and compare the reported performance measures and validation strategies across studies in order to summarize the quantitative strength of evidence within comparable subgroups, emphasizing metrics that reflect decision usefulness under class imbalance such as recall for high-severity outcomes, F1-score, AUC, and calibration indicators. The review also seeks to document dataset characteristics that influence model reliability and generalization, including sample size, event prevalence, label taxonomy, missing-data patterns, temporal coverage, and contextual factors such as industrial sub-sector, project or plant type, and geographic setting. In addition, the review aims to assess how studies operationalize "injury reduction relevance" by examining whether predictions are tied to actionable control measures, whether risk proxies such as unsafe-condition detection are validated against injury outcomes or credible safety performance indicators, and whether evaluation designs reflect realistic deployment constraints through temporal testing, site-held-out validation, or external dataset testing. Finally, the review intends to synthesize how methodological choices—feature engineering, imbalance handling, interpretability methods, and robustness checks—affect reported performance and comparability, producing a structured evidence map that distinguishes injury prediction models from hazard-state detection models while preserving quantitative comparability within each category.

LITERATURE REVIEW

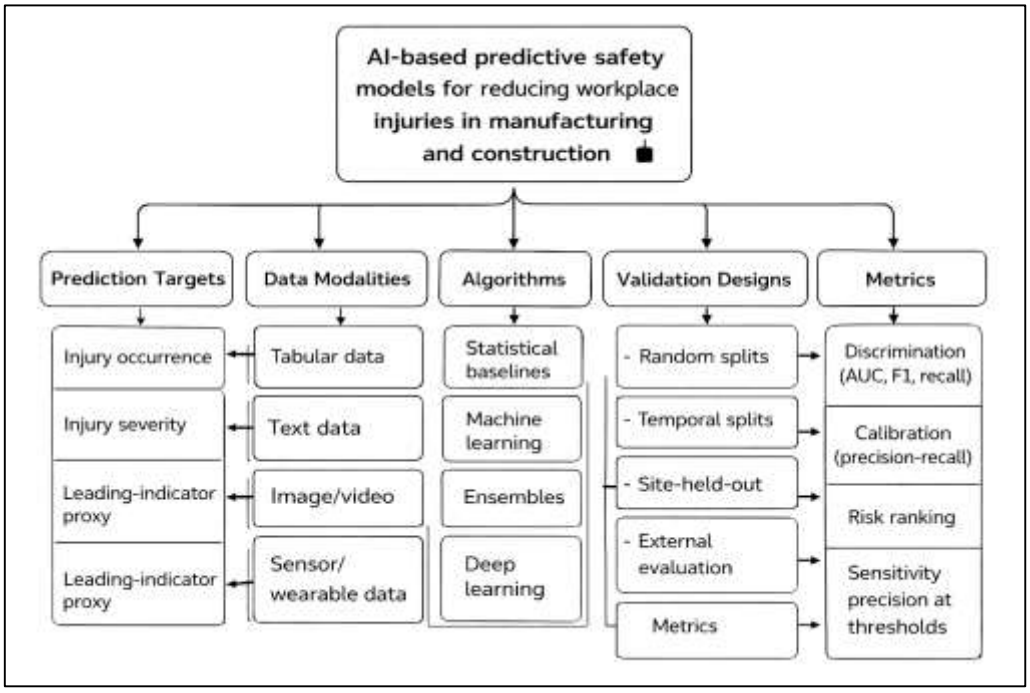
The literature review for a systematic review on artificial intelligence (AI)–based predictive safety models must establish a structured, evidence-mapping foundation that aligns manufacturing and construction safety research with the quantitative logic of prediction, validation, and measurable outcomes (Toronto, 2020). Within occupational safety science, "predictive safety" is operationalized through models that estimate injury likelihood, classify injury severity, or infer high-risk conditions and behaviors that are statistically associated with subsequent injury events. Because manufacturing and construction differ in work-process stability, hazard exposure patterns, and data availability, the literature base is methodologically diverse: manufacturing studies commonly rely on structured logs (incident databases, production and maintenance records, shift schedules), while construction studies frequently integrate unstructured narratives, images, and dynamic site sensing due to rapidly changing work zones and concurrent operations (Dmitrienko et al., 2020). The literature review therefore must do more than summarize findings; it must classify evidence according to prediction target (injury occurrence vs severity vs leading-indicator proxies), data modality (tabular, text, vision, sensor), modeling family (classical ML, ensemble methods, deep learning), and evaluation design (random split vs temporal split vs site-held-out vs external validation). This section synthesizes what existing studies collectively reveal about model performance, reliability, and comparability by extracting and organizing quantitative indicators such as dataset size, event prevalence, class imbalance ratios, performance metrics (AUC, F1, recall for severe events, calibration), and deployment-relevant constraints (input availability, prediction horizon, and action unit). By structuring the literature in this way, the review can identify which combinations of data sources, algorithms, and validation methods are most frequently used, which model objectives are most studied in each sector, and where quantitative evidence clusters sufficiently to support meaningful subgroup synthesis (Gulcin, 2020).

AI-Driven Workplace Injury Prediction

Prediction targets are operationally defined as the dependent outcomes the model attempts to estimate. Injury occurrence prediction refers to models trained to estimate whether an injury event will occur within a defined unit of analysis, which may be framed as a binary classification (injury vs no injury) or as a probabilistic risk score anchored to a time window (for example, a shift, a day, or a week). Injury severity prediction refers to models trained to classify or rank the consequence level of an incident, typically expressed as multiclass categories that represent medical and operational burden, including outcomes that imply time loss or permanent harm. Leading-indicator proxy prediction refers to models

trained to estimate measurable precursor states such as noncompliance signals, unsafe condition detection, hazardous interaction configurations, or near-miss likelihood, where the proxy is treated as injury-reduction relevant when it is empirically tied to safety performance indicators or incident outcomes within the study design (Aggarwal et al., 2021). Data modalities are defined as the structure and source of predictor variables. Tabular modality includes coded records such as incident logs, audits, training histories, maintenance indicators, production variables, staffing patterns, and environmental measures. Text modality includes unstructured narratives from incident descriptions, safety observations, inspection comments, and corrective-action notes, where natural language is transformed into features. Image and video modality includes visual sources from cameras and site imagery used to detect workers, equipment, protective gear, and hazardous conditions. Sensor and wearable modality includes time-stamped streams from proximity devices, location tags, inertial sensors, and physiological measures used to quantify exposure and interaction patterns. Algorithm families are categorized to support interpretation of performance differences: baseline statistical models are treated as comparators; machine learning classifiers include traditional supervised learners; ensembles include tree-based aggregation and boosting; deep learning includes neural architectures aligned to text, vision, and time-series data. Validation designs are defined by the mechanism used to test generalization: random splits evaluate internal discrimination; temporal splits assess stability across time; site-held-out designs test transfer across worksites or projects; external validation tests performance across independent datasets (C. Huang et al., 2022). Metrics are defined as the quantitative measures used to summarize model performance, emphasizing discrimination and decision relevance under class imbalance through measures such as AUC, F1, recall or sensitivity, precision, probability calibration, and precision–recall area measures. These operational definitions allow the literature to be grouped into comparable subsets and synthesized without mixing fundamentally different outcomes, data-generating conditions, and evaluation standards.

Figure 3: AI-Driven Workplace Injury Prediction



Across the empirical literature, quantitative comparability is repeatedly shaped by dataset characteristics and by the realism of validation strategies, which determines whether reported model performance reflects operational utility or only in-sample pattern recognition (Zhang et al., 2022). A consistent pattern across many studies is that injury data are often imbalanced, meaning severe outcomes and recordable injuries occur far less frequently than non-injury observations or low-severity cases, and this imbalance affects the meaning of commonly reported metrics. The literature also shows

that predictive models trained on a single organization, site, or project can achieve high internal discrimination while failing to transfer when reporting practices, work processes, or hazard controls differ. As a result, the review emphasizes dataset descriptors that influence interpretability: total sample size, the number of injury events, event prevalence, class distribution across severity categories, missing-data proportions, duplication risk, and the stability of feature definitions across time. Many studies use structured incident databases as the primary substrate for modeling, yet incident logs frequently contain inconsistent coding and narrative variability, and they may lack precise exposure denominators that would support more rigorous risk-rate modeling. Text-based studies often demonstrate the usefulness of unstructured narratives for capturing contextual precursors, but they also reveal sensitivity to vocabulary differences, report-writing conventions, and label mapping decisions that can shift class boundaries (Razavykia et al., 2020). Vision-based studies often measure detection quality using frame-level or object-level metrics that quantify correct identification of protective equipment or hazardous configurations, yet these metrics do not automatically translate into injury prediction unless the detected states are linked to incident outcomes or validated safety indicators. Sensor and wearable studies frequently provide granular exposure measures and high-frequency interaction signals, but their datasets are often narrower in scope, and their performance claims may depend on controlled deployment conditions or limited site diversity. For these reasons, validation design is treated as a primary credibility indicator: random split validation is common but vulnerable to leakage when the same site patterns appear in both training and testing; temporal validation better reflects real-world deployment by separating earlier and later periods; site-held-out validation directly tests whether models generalize across projects or plants; and external validation tests whether the model transfers across organizations or regions. The literature also indicates that decision thresholds and alert rates matter operationally, so studies that report sensitivity and precision at specific thresholds provide stronger evidence for safety decision contexts than studies that report only global discrimination measures (Sarker et al., 2020). This systematic structure ensures that findings from numerous studies can be compared on common quantitative ground without forcing equivalence between incompatible tasks.

Finally, the taxonomy supports synthesis by clarifying how different model objectives relate to injury reduction in manufacturing and construction through measurable outputs and quantifiable evaluation, rather than through general claims. In injury occurrence prediction research, studies often frame risk estimation at the level of worker-shift, crew-day, or site-week, using structured operational inputs to generate probability scores that can be ranked for prioritization (Ozturk, 2021). In severity prediction research, models frequently focus on classifying high-burden outcomes, because identifying a smaller subset of high-consequence cases can have disproportionate safety value when resources are limited. In leading-indicator proxy research, predictive models estimate frequent, measurable risk states—such as unsafe proximity patterns, missing protective gear, or noncompliance signals—that can be tracked continuously and used as quantitative indicators of safety conditions. The literature indicates that each target type requires different evidence standards: injury outcomes require careful handling of rare-event imbalance, while proxy outcomes require credible linkage to injury-related performance measures. This distinction becomes especially important across sectors because manufacturing settings often support richer structured inputs and stable process signals, while construction settings often require multi-modal fusion of narratives, imagery, and dynamic exposure measures (Boboc et al., 2022). Algorithm choices in the literature reflect these constraints: structured tabular datasets are frequently modeled with ensemble learners and classical classifiers due to strong performance under mixed features; unstructured text is often modeled through language-based feature representations and deep learning approaches; imagery is typically modeled using deep visual detectors; and time-series sensor streams are modeled using temporal learners designed for sequential dependence. Metrics then function as the common language for synthesis across these heterogeneous settings, but only when interpreted within the correct task class: AUC and related discrimination measures support ranking performance, recall and sensitivity reflect capture of high-risk cases, precision reflects false-alarm control, and calibration reflects whether predicted probabilities can be treated as trustworthy risk estimates (Cagnano et al., 2020). The review therefore treats the literature review section as an evidence map built on operational definitions and quantitative comparability rules, enabling systematic

grouping and synthesis of results across many studies while keeping the focus on measurable prediction objectives within manufacturing and construction safety contexts.

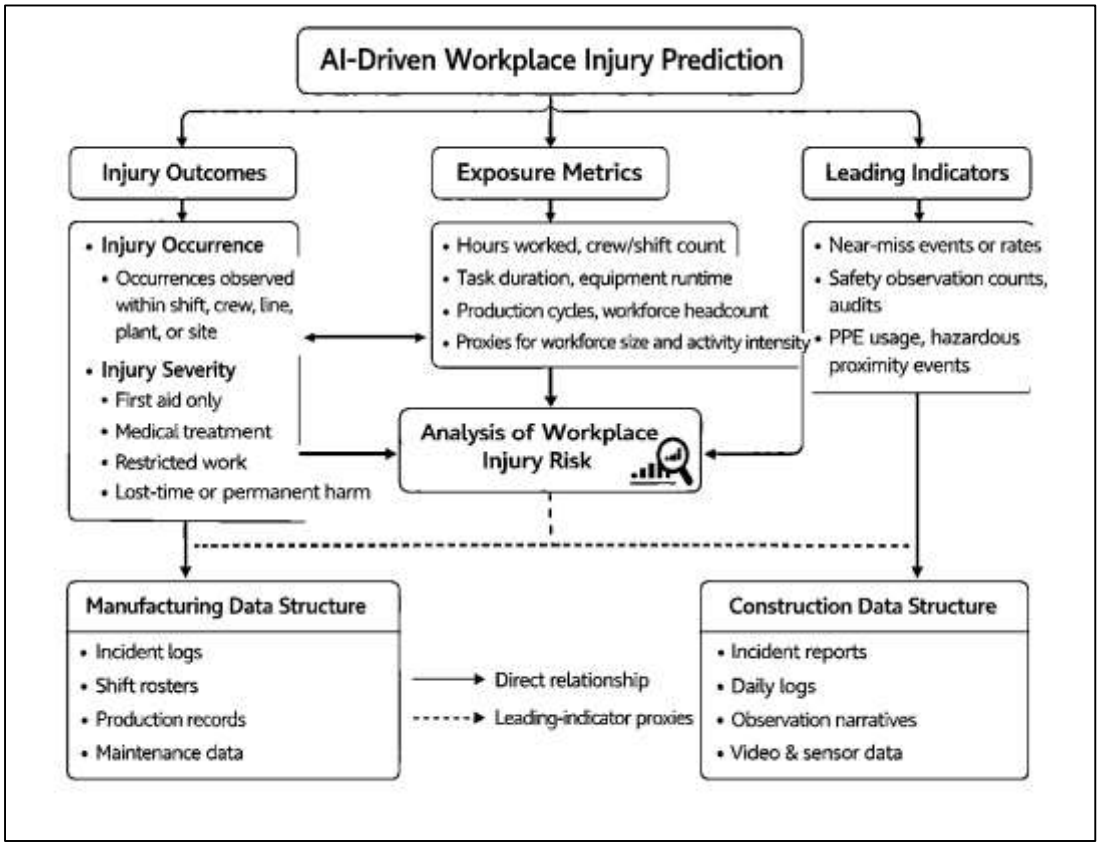
Safety Outcome Constructs

Workplace injury outcomes in manufacturing and construction are typically defined in ways that support consistent counting, classification, and severity grading for quantitative analysis. In the predictive safety literature, injury occurrence is most often operationalized as an event outcome observed within a specified unit of time or unit of work, enabling a model to learn the difference between periods with and without recorded injuries (Bayramova et al., 2023). This operationalization appears in binary formulations where the outcome indicates whether an injury occurred during a worker-shift, crew-day, line-week, or project-month, as well as in count-based formulations where the outcome represents the number of injuries recorded within a defined exposure window. Studies that use occurrence outcomes frequently draw from administrative injury logs, incident reporting systems, and safety management databases because these sources provide time-stamped records and standardized fields that can be aligned to predictors such as job type, task category, equipment involved, or location. In addition to occurrence, severity categories are widely used because the safety burden of occupational harm is not uniform across events. Severity is commonly treated as an ordinal or multiclass outcome aligned to medical and operational consequences, including cases that require first aid only, medical treatment beyond first aid, restricted work, lost-time cases, permanent impairment, and fatal outcomes (van Nunen et al., 2022). Severity labels allow predictive models to focus on high-consequence events that impose larger costs and higher human harm, while still accommodating the full incident distribution. Exposure metrics form a third foundational construct because injury probability depends on opportunities for exposure to hazards. The literature uses exposure as the denominator that helps interpret counts and probabilities, including hours worked, shift duration, task time, equipment runtime, production cycles, and workforce headcount. When exposure is measured and linked to outcomes, models can represent risk in a way that better reflects differences between high-activity and low-activity periods. Even when exposure is not directly modeled as a denominator, it is often included as a predictor because overtime, extended shifts, and high-intensity work periods are treated as measurable correlates of injury occurrence and severity. Across these outcome definitions, a persistent theme is that injury data are shaped by reporting behavior, coding practices, and organizational procedures, which affects label reliability (Abdul, 2023; Hammad & Mohiul, 2023; May et al., 2022). As a result, the literature emphasizes careful mapping of outcome categories, consistent time windows, and transparent rules for aggregating events into model-ready outcomes so that predictive performance can be interpreted in comparable ways across studies that differ in site type, dataset size, and event prevalence.

Manufacturing and construction differ in injury data structure because they differ in workflow stability, task repetition, and the granularity of routine operational records, and these structural differences shape how injury outcomes and exposures are represented in quantitative models (Hasan & Waladur, 2023; Ta et al., 2020). In manufacturing, many studies leverage the high density of structured data produced by stable production systems, including shift rosters, production throughput logs, quality records, maintenance events, machine condition indicators, and standardized incident forms. This environment supports modeling pipelines where predictors are consistently recorded at regular intervals, enabling alignment between predictors and injury outcomes at the shift, line, department, or plant level (Rifat & Rebeka, 2023). Stable processes also support repeated observation units, which increases sample size and strengthens statistical learning when outcomes are rare . However, structured density does not eliminate measurement challenges: injury events may still be underreported, near misses may not be consistently logged, and coding fields may be incomplete, requiring preprocessing decisions that directly influence model comparability. In construction, the literature describes a higher fraction of unstructured and semi-structured data because worksites are dynamic and project conditions change rapidly (Leso et al., 2023; Masud & Hossain, 2024; Md & Sai Praveen, 2024). Incident narratives, safety observation notes, inspection comments, corrective-action descriptions, and daily logs become important sources of context because they record transient risk factors such as changing work zones, simultaneous operations, equipment movement patterns, weather conditions, and subcontractor coordination (Nahid & Bhuya, 2024; Newaz & Jahidul, 2024).

Construction studies therefore often require additional steps to convert language-based and observational material into quantifiable inputs, and outcomes may be defined at varying levels such as activity, crew, location zone, or project phase rather than at a stable production line. This unstructured emphasis also appears in the use of images and video, where site cameras or mobile devices provide visual records of safety states that do not exist in standardized numeric fields. The data structure differences across the two sectors affect how exposure is captured: manufacturing often has direct measures such as runtime and throughput, whereas construction exposure is frequently approximated through workforce counts, task durations, or schedule-based measures extracted from site records (Hidayati et al., 2020; Akbar, 2024; Rabiul & Alam, 2024). These sector-specific structures create a comparability problem addressed by systematic reviews through explicit categorization: a model trained on manufacturing's consistent structured logs is not directly comparable to a model trained on construction narratives and visual streams unless the review isolates the prediction target, defines how the unit of analysis is constructed, and records the dataset's granularity and labeling rules. The literature therefore treats "data structure" as a quantitative moderator that influences reported performance, because feature completeness, label stability, and temporal alignment are easier to achieve in structured environments and more variable in dynamic project contexts. Within this framing, systematic synthesis requires documenting not only what outcomes were predicted but also how the underlying observation units were formed and how predictors were captured in each sector.

Figure 4: AI-Based Predictive Safety Outcomes



Leading indicators are used in the predictive safety literature to represent measurable precursors and risk states that occur more frequently than injuries and can be quantified continuously, allowing models to learn patterns of elevated risk even when injury events are rare (Blut & Wang, 2020; Hammad & Hossain, 2025; Azam & Amin, 2024). Three leading-indicator families are commonly emphasized: near-miss frequency, safety observation counts and audit nonconformance, and compliance-related measures such as protective equipment usage and hazardous proximity events. Near misses are typically defined as unplanned events that did not result in injury but had the potential to do so, and they are modeled as counts or rates within a unit such as a shift, week, or project phase. Safety

observations and audit nonconformance are similarly treated as countable events in which unsafe conditions, unsafe acts, or procedural deviations are recorded through inspections, behavioral observations, or audit processes (Mosheur, 2025). These measures are attractive because they create larger datasets than injury logs and can support more stable learning, yet they also introduce comparability issues because observation intensity varies with inspection frequency, supervisor practices, and organizational reporting culture. The literature therefore treats observation processes as part of the measurement construct, meaning that comparisons across studies require attention to how observation data were collected, who recorded them, and whether the data reflect systematic sampling or opportunistic reporting (Zaheda, 2025a, 2025b). A second major proxy category is personal protective equipment compliance, which appears both in structured observation checklists and in computer-vision studies that quantify compliance from images or video (Hassan et al., 2023). PPE compliance is typically expressed as a proportion of compliant observations within a defined period, and it is used as a measurable indicator of safety behavior and control effectiveness. The third category is proximity-related hazard indicators, which are derived from location tracking, proximity sensors, or equipment interaction monitoring and represent measurable exposure to hazardous interactions between workers and moving equipment or hazardous zones. These proximity indicators are treated as quantitative proxies for risk intensity because they capture how long and how often workers operate within hazardous spatial relationships, which is difficult to infer from incident logs alone. Across leading indicators, the literature positions proxies as injury-reduction relevant when they represent controllable risk states and when studies demonstrate meaningful association with incident outcomes, severity patterns, or validated safety performance measures. Because leading indicators differ in frequency, reliability, and susceptibility to reporting bias, systematic synthesis requires categorizing them by construct type and documenting their operational definitions, measurement intervals, and data completeness (O'Donovan & McAuliffe, 2020). The evidence base also indicates that proxy outcomes are not interchangeable with injury outcomes: they provide broader sampling of risk states but require careful interpretation because they reflect both underlying risk and the intensity of observation systems. This is why systematic reviews emphasize clear definitions and transparent measurement rules for leading indicators so that performance comparisons across predictive models remain grounded in comparable constructs rather than in mixed measurement processes.

Bringing injury outcomes, sector-specific data structure, and leading indicators together, the literature review framework for quantitative safety outcome constructs emphasizes consistent operational definitions and transparent measurement decisions as the basis for credible comparison across predictive safety studies (Jiang et al., 2021). Injury occurrence and injury severity outcomes provide direct measures of harm and burden, yet they are often constrained by rarity, reporting variation, and inconsistent coding, which can complicate model training and inflate apparent performance if validation is not robust. Exposure metrics serve as the interpretive bridge that links injuries to opportunities for risk, allowing models to represent variation in workload, time at risk, and equipment interaction intensity. Manufacturing datasets frequently enable clearer exposure linkage due to stable production rhythms and routine logging, while construction datasets often require proxies and approximations because exposure varies with schedule changes, subcontractor presence, and shifting site conditions. Leading indicators complement injury outcomes by producing higher-frequency measurements of risk states, including near-miss logs, observation systems, audit findings, PPE compliance measures, and proximity-based exposure indicators. These proxies expand the evidentiary base for predictive modeling but also broaden the measurement problem, since leading indicators are shaped by surveillance intensity, reporting incentives, and observational coverage. In response, the predictive safety literature increasingly treats outcome constructs as multi-layered: injury outcomes represent realized harm; severity represents consequence; exposure represents opportunity for harm; and leading indicators represent measurable precursors and control-relevant states (Rockström et al., 2023). A systematic review that synthesizes this literature must therefore record how each study defines its outcome construct, how observation units are formed, and how exposure and proxy measurements are captured and aligned to model inputs. Comparability depends on identifying whether a study predicts injuries directly, predicts severity among recorded incidents, or predicts precursor states that are positioned as risk signals. It also depends on documenting whether outcomes and indicators are

measured at the worker level, crew level, line level, or site level, because the unit of analysis affects event prevalence, label stability, and the meaning of performance metrics. Finally, the literature emphasizes that the quality of outcome constructs is inseparable from the quality of the data-generating system that produced them, meaning that transparent reporting of data sources, labeling rules, missingness, and observation processes is necessary to interpret predictive performance across manufacturing and construction contexts (Means et al., 2020). This outcome-construct framing supports systematic categorization and subgroup synthesis without conflating fundamentally different targets, enabling the review to compare like with like in a quantitatively defensible way.

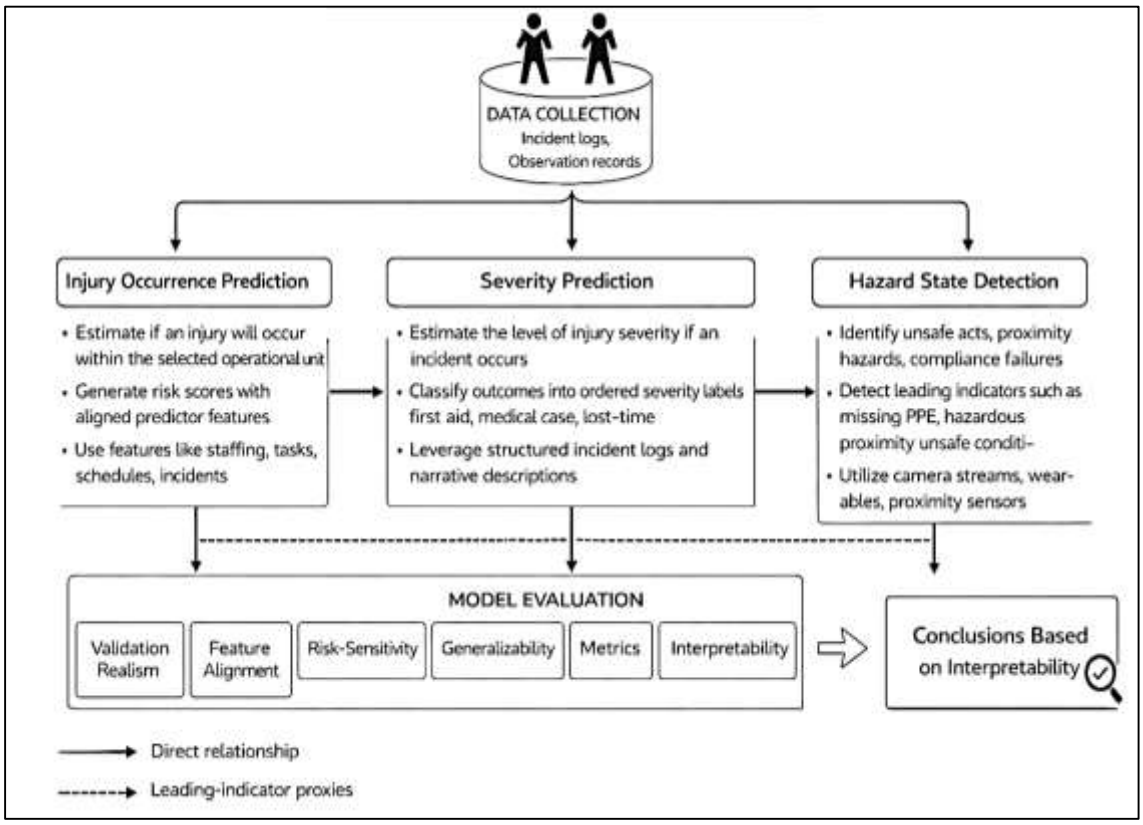
Predictive Safety Modeling

In the predictive safety literature, injury occurrence prediction models are defined by their aim to estimate whether an injury event will occur within a specified operational unit, and this unit-of-prediction choice shapes the structure of inputs, the meaning of model outputs, and the interpretability of results (Mehdizadeh et al., 2020). Studies commonly align prediction units to the cadence of safety decision-making, such as worker-shift, crew-day, equipment-zone, or site-week units, because these align with staffing assignments, production planning, and routine safety oversight. Worker-shift models typically integrate worker attributes, task assignments, shift length, overtime exposure, training records, and recent incident history to produce a risk estimate that can be used for prioritizing supervision or targeted briefings within a single shift. Crew-day and site-week models often aggregate individual-level inputs into group-level indicators such as crew composition, subcontractor mix, work package type, schedule pressure proxies, and environmental conditions, enabling a broader view of risk distribution across concurrent operations. Equipment-zone models foreground spatial exposure and interaction risk, often combining location-based features, equipment movement indicators, and zone-specific hazard characteristics to estimate elevated risk states tied to particular work areas. Across these variants, the literature consistently treats model outputs as decision-support signals, frequently expressed as a risk score or hazard category that supports ranking of operational units by relative risk (Bartulović & Steiner, 2023). This output framing is particularly common when injury outcomes are rare, because ranking performance and risk stratification are more stable than exact numeric probability estimation in sparse-event contexts. Even when studies report probabilistic outputs, interpretive emphasis often remains on how effectively the model separates higher-risk from lower-risk units under class imbalance. A recurring methodological theme is that the unit of prediction must match the unit of control: a model that predicts risk at the site-week level supports managerial resource allocation and inspection planning, while a worker-shift model is more relevant to frontline supervision and task-level planning. The literature also shows that model inputs must be aligned to the prediction window to avoid leakage, meaning that the features must reflect information available before the event window closes. For example, if the prediction unit is a shift, features should be measurable at or before shift start, or at defined time checkpoints, rather than extracted from end-of-shift reports that embed post hoc incident information. This alignment is critical because safety datasets often contain time-stamped narratives and corrective action notes that can inadvertently reveal outcome information. Consequently, injury occurrence models are frequently discussed in relation to validation realism, including whether the data split respects time order, whether entire sites or projects are held out, and whether model performance holds across operational contexts with different reporting practices (Hu et al., 2020). Within manufacturing and construction, the occurrence-prediction literature therefore emphasizes the practical mechanics of constructing comparable prediction units, generating outputs that support triage, and evaluating models in ways that reflect the dynamic and heterogeneous settings where injury risk is managed.

A second major objective class in predictive safety research is severity prediction and classification, where models are designed to estimate the consequence level of an injury event or to classify incident outcomes into ordered or multi-category severity labels. Severity modeling appears in two primary forms (Elmaz et al., 2021). In one form, studies model severity conditional on an incident having occurred, using incident descriptors, contextual features, and narrative content to classify whether the outcome involved first aid only, medical treatment, restricted duty, lost time, permanent impairment, or fatality. In another form, studies integrate severity into broader prediction pipelines by estimating the likelihood of high-severity outcomes within operational units, effectively combining occurrence

and severity into a burden-oriented risk signal. The literature emphasizes severity modeling because high-severity injuries represent a smaller share of incidents but a disproportionately large share of human and economic costs, making them a distinct analytic target. From a quantitative standpoint, severity modeling is often harder than occurrence prediction because label distributions are more imbalanced and because severity can be influenced by contextual factors that are not consistently recorded, such as immediate response quality, reporting timeliness, and local medical thresholds. Studies therefore frequently treat rare-event modeling as a central design constraint, using strategies such as class weighting, resampling, and threshold-focused evaluation to improve sensitivity to high-severity outcomes (Viceconti et al., 2021).

Figure 5: AI Predictive Safety Model Framework



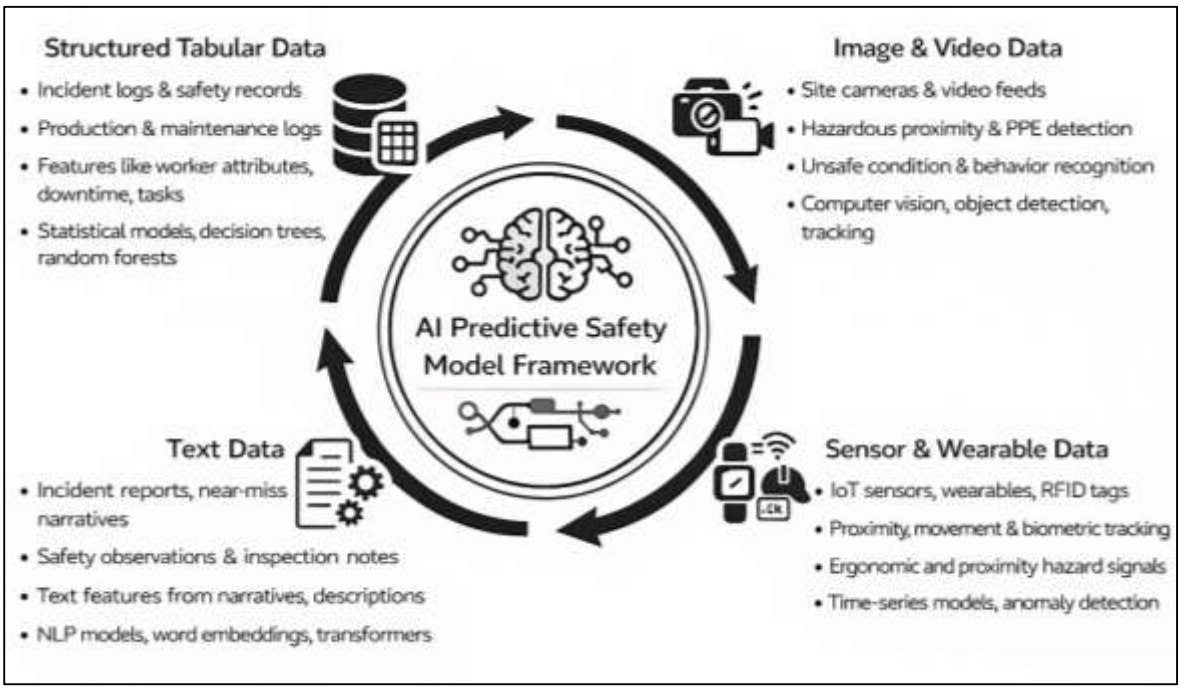
A consistent observation across severity studies is that performance varies by severity class and that global accuracy metrics can obscure poor detection of rare but critical categories. Consequently, many studies emphasize class-specific performance reporting, including how well models identify high-severity cases relative to low-severity cases. Severity prediction is also linked to the data modality used. Structured incident logs often provide categorical fields for injury type, body part, equipment involved, and coded cause categories, which support classical supervised learning. At the same time, unstructured narrative fields often contain richer contextual detail about circumstances, sequence of events, and immediate consequences, making text-driven features particularly prominent in severity modeling research. This has led to a substantial body of work in which narratives are transformed into predictive representations and used to classify severity outcomes, often alongside structured fields. In construction contexts, severity modeling also intersects with project-level variability: the same incident type can yield different severity outcomes depending on height, energy source, work posture, or surrounding constraints, which complicates generalization across sites. In manufacturing contexts, stable processes can support larger sample sizes, but severe outcomes remain rare, and severity may cluster around specific equipment classes or tasks, raising the need for validation that tests generalization beyond a single plant or production line (Zhang & Mahadevan, 2020). Across both sectors, the literature frames severity prediction as a triage and learning tool that can standardize

severity coding, support prioritization of investigations, and identify patterns in high-consequence incidents, while also highlighting that severity labels reflect both injury harm and reporting systems. This places emphasis on transparent severity taxonomy definitions, consistent label mapping, and evaluation designs that prevent optimistic estimates arising from duplicated narratives or repeated site-specific wording patterns.

Data Modalities Used in AI Predictive Safety Models

Structured tabular data represent the most established modality for AI predictive safety models, particularly in manufacturing environments where stable processes and standardized reporting systems generate consistent records over long periods. In this literature, tabular inputs most often originate from incident logs and safety management databases that encode event attributes such as injury type, body part affected, task category, equipment involved, and coded causal or contributing factors (Tselentis et al., 2023). These datasets are frequently merged with production and maintenance systems that record downtime episodes, repair frequency, preventive maintenance compliance, backlog measures, and machine condition indicators, allowing models to represent operational strain and equipment reliability as measurable correlates of injury risk. Workforce attributes are also common structured inputs, including tenure, training completion, role classification, overtime exposure, shift patterns, and staffing levels, because they provide quantifiable proxies for experience, fatigue, and workload distribution. Within systematic evidence mapping, tabular studies are primarily comparable when they clearly report dataset scale and event structure, since predictive results are strongly moderated by sample size, injury event counts, and the base rate of injuries in the observation window. Many manufacturing datasets are large in record count but sparse in positive events, making imbalance a recurring characteristic and shaping what performance metrics are meaningful (Cai, 2020).

Figure 6: AI Predictive Safety Data Framework



In addition, missingness is a prominent extraction item because incident logs and operational systems often contain incomplete fields, inconsistent coding, and variable reporting quality across departments, which can materially change model inputs and downstream performance. As a result, the tabular-data literature frequently emphasizes preprocessing decisions such as imputation strategies, feature selection, categorical encoding, aggregation rules for constructing worker-shift or line-week units, and temporal alignment between predictors and outcomes. These choices affect whether a model captures true risk signals or artifacts of recordkeeping. The reviewed studies often present tabular modeling as the most deployable path for predictive safety because the data sources already exist in many

organizations and require minimal additional instrumentation, yet the evidence also shows that tabular datasets can encode structural bias through underreporting, inconsistent event taxonomy use, and differences in safety culture across sites. Accordingly, systematic review extraction from tabular studies typically records not only the variables used but also the definitional rules that convert raw records into outcomes and observation units, because cross-study comparability depends on whether injuries are labeled consistently and whether the exposure window matches the operational decision context (Lee et al., 2024). This makes structured tabular research a backbone of AI predictive safety modeling, while also making transparent reporting of dataset characteristics an essential requirement for comparing findings across manufacturing and construction contexts.

Text data form a second major modality in AI predictive safety models, capturing information that structured logs often cannot represent, including contextual sequences, narrative descriptions of conditions, and nuanced descriptions of causal factors embedded in incident reports, near-miss narratives, safety observations, inspection notes, and corrective-action documentation (Salhab et al., 2024). In the literature, text-driven predictive models are motivated by the observation that narrative fields frequently contain details about how events unfolded, what immediate precursors were present, and what environmental or organizational conditions contributed to risk, even when structured codes are missing or overly coarse. Text-based pipelines typically follow a sequence of steps that transform unstructured language into model-ready representations, including cleaning and normalization, tokenization, and representation through either count-based vectorization, embedding-based features, or deep language encodings. Across studies, a key comparability element is the size and nature of the feature representation, because vocabulary breadth, embedding dimensionality, and the use of domain-specific language modeling can substantially affect performance. Another essential extraction item is how labels are mapped from text-linked records to outcomes such as severity category, incident type, or contributing-factor class (Charoenpitaks et al., 2024).

Label mapping is particularly consequential because the same narrative can be associated with multiple classification targets and because organizational taxonomies differ across sites and sectors. The literature also documents that narrative length and narrative completeness vary widely by reporter, organizational practice, and reporting platform, making distributional descriptors of narrative length and class counts informative for interpreting model performance. Text modeling in construction receives special attention because unstructured documentation is often central to project safety oversight, and site conditions change rapidly in ways that are more likely to be recorded in notes than in standardized numeric fields. At the same time, manufacturing also benefits from text analysis when incident narratives and maintenance notes contain signals about recurring hazards, procedural nonconformance, or equipment behaviors that precede injuries. The evidence base indicates that text models can perform well in classifying severity or categorizing incident types, especially when narratives are plentiful and label definitions are stable, yet performance can degrade when narrative templates differ across sites or when terminology varies across regions and trade groups. For systematic synthesis, text studies are therefore commonly grouped by outcome target and reporting context, and extraction emphasizes narrative corpus size, the number of labeled cases per category, and whether validation designs control for leakage arising from repeated wording patterns (Xu et al., 2024). Overall, text modalities broaden the predictive safety evidence base by converting qualitative descriptions into quantitative signals, but comparability depends on transparent reporting of language preprocessing, feature construction, and label taxonomy alignment.

Sensor, wearable, and IoT data form a fourth modality that bridges manufacturing and construction and is distinctive because it produces time-stamped streams that directly measure exposure intensity, interaction patterns, movement dynamics, and physiological or biomechanical proxies related to fatigue and ergonomic risk (Khowaja et al., 2022). In construction, proximity sensors and location tags are commonly used to quantify interactions between workers-on-foot and heavy equipment, producing measurable indicators of hazardous proximity events and exposure duration within dangerous zones. In manufacturing, inertial sensors and wearable platforms are often used to monitor posture, repetition, force proxies, and movement variability, supporting prediction tasks related to ergonomic risk states and unsafe motion patterns that can precede injuries. Physiological measures, where used, can represent strain proxies that relate to heat stress or fatigue-related risk. The literature frequently

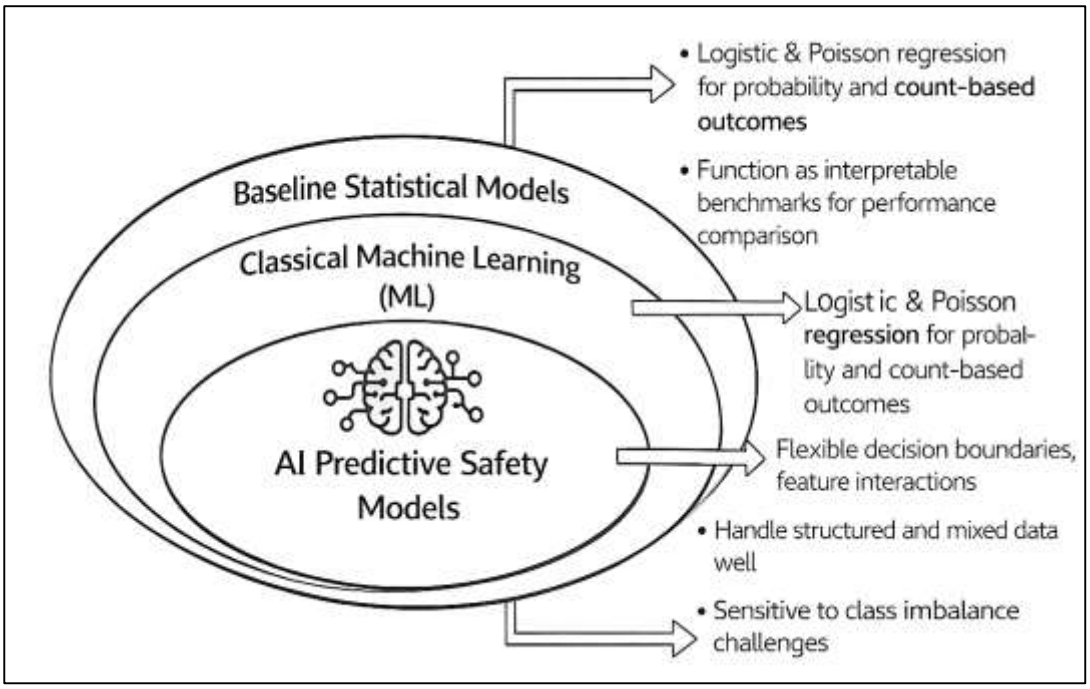
operationalizes these data through observation windows that aggregate streams into derived features such as proximity event frequency, dwell time in hazardous zones, acceleration-based movement descriptors, posture classification counts, or workload indicators. Comparability in this modality depends on reporting the sampling characteristics and the windowing strategy used to construct prediction inputs, because sampling frequency and observation window length directly shape what patterns the model can learn (Serradilla et al., 2022). Another key extraction item is the number of derived features and how they are engineered, since time-series modeling can range from feature-based classical learners to sequence-based deep learning approaches depending on data volume and labeling density. Performance reporting often focuses on the ability to detect or classify risk states within defined windows, with emphasis on sensitivity at decision thresholds because operational deployment requires control of missed detections and false alarms. Sensor and wearable studies also face distinctive data quality issues, including signal loss, calibration drift, device noncompliance, coverage gaps, and environmental interference, all of which can bias both training data and evaluation results. As with vision-based models, injury-reduction relevance hinges on measurable linkage: proximity events and ergonomic risk states are treated as predictive safety outputs when they correspond to validated hazard mechanisms and when their measurement supports actionable controls such as exclusion zone management, equipment routing, task rotation, or targeted training. For systematic synthesis, sensor/wearable studies are therefore grouped by hazard construct measured, sensor type, labeling method, and whether the outcome is an injury event, a severity category, or a risk proxy (K. Huang et al., 2022). These studies contribute uniquely to predictive safety by providing objective exposure measures that are otherwise difficult to obtain from logs and narratives, while requiring clear reporting of sampling, windowing, feature derivation, and evaluation protocols for cross-study comparability.

Algorithm Families and Model Architectures

Baseline statistical models form the methodological anchor in many AI-based predictive safety studies because they provide interpretable comparators and establish whether more complex algorithms deliver measurable value beyond conventional inference. In this body of literature, logistic regression is widely used for injury occurrence outcomes defined as binary events at operational units such as worker-shift, crew-day, or site-week, because it produces stable estimates under modest data sizes and supports straightforward inclusion of structured predictors such as overtime exposure, tenure, training completion, task class, and equipment involvement (Wang et al., 2023). Count-based modeling approaches, including Poisson and negative binomial specifications, appear when outcomes are defined as incident counts within exposure windows and when researchers aim to align predictions with event frequency rather than simple occurrence. These models are particularly relevant when injury events are aggregated at line-week, department-month, or project-phase levels and when exposure-related predictors such as hours worked or runtime are treated as key quantitative descriptors. Baseline models are repeatedly positioned in the literature as quality benchmarks because they encourage transparent feature specification, support examination of confounding and collinearity, and enable clearer diagnosis of whether predictive gains are attributable to nonlinear learning or to differences in variable handling and preprocessing. In systematic evidence mapping, baseline studies are frequently used to evaluate whether performance improvements reported by machine learning are robust to differences in data preparation and outcome construction. They also provide a consistent interpretive reference when datasets are sparse, when high-severity outcomes are rare, or when labels are noisy due to reporting variability (Aslan & Yilmaz, 2021). The baseline literature further highlights that predictive performance depends not only on the algorithm but on how predictors are engineered and aligned temporally to the prediction window. For example, injury occurrence modeling can be inflated by features that inadvertently encode post-event information, and baseline modeling frameworks often make these risks easier to detect because feature definitions are typically documented in a more explicit manner. Another recurring methodological contribution of baseline modeling is its role in calibration and probability interpretation, since many safety decision contexts require probability-like outputs that can be compared across units and thresholds. When safety studies report results using discrimination measures alone, baseline models still provide a useful reference for how well probability estimates align with observed event frequencies under class imbalance. Across manufacturing and construction, baseline statistical modeling therefore functions as both a practical

method and a methodological control: it sets a minimum evidentiary standard for predictive claims and clarifies whether complex model families contribute incremental accuracy, improved sensitivity to rare events, or better generalization across sites and time windows (Gasparetto et al., 2022).

Figure 7: AI Predictive Safety Model Hierarchy



Classical machine learning models extend the predictive safety literature beyond parametric baselines by introducing flexible decision boundaries and feature interactions while still remaining relatively tractable in terms of implementation and interpretation. Decision trees are frequently used because they capture nonlinear splits and rule-like logic that safety practitioners can often interpret as actionable patterns, especially when predictors include categorical codes from incident logs, job classifications, hazard categories, and compliance indicators (Bienvenido-Huertas et al., 2020). k-nearest neighbors approaches appear less consistently but remain conceptually relevant in safety prediction because they operationalize similarity-based reasoning, classifying new observations by proximity to historical cases in a feature space constructed from worker attributes, task descriptors, or environmental conditions. Support vector machines are commonly evaluated in comparative studies because they can perform well in high-dimensional feature spaces, including settings where text-derived vectors are used to classify severity, incident type, or causal categories. Naïve Bayes remains a frequent baseline in text-heavy safety studies because it is robust, simple, and often competitive when narrative fields are transformed into sparse representations. Across these classical models, the literature emphasizes that comparability and performance depend heavily on extraction items that are sometimes inconsistently reported, including hyperparameter specification, feature scaling methods, and strategies for managing class imbalance. Safety datasets often contain rare outcomes, especially when focusing on high-severity injuries, and classical models can be sensitive to imbalance when default decision thresholds are used or when class weights are not explicitly managed (Papa et al., 2024). This makes class-weighting, resampling approaches, and threshold tuning important elements for systematic extraction, because they directly influence recall for high-risk categories and false-alarm rates in operational use. Feature scaling also becomes important for distance-based learners and margin-based methods; without clear documentation of scaling, performance comparisons can be misleading. Another recurring theme in safety modeling research is that classical algorithms can behave differently depending on the unit of analysis: a worker-shift dataset with many repeated workers introduces clustering that can inflate performance under random splitting, while a site-held-out design can reveal whether a model is learning general risk mechanisms or site-specific coding

patterns. In construction and manufacturing safety studies that include mixed structured and derived features, classical models are often used as strong baselines against which ensembles and deep learning are compared, not because they are universally superior, but because they help clarify the marginal value of complexity under real-world data constraints. Their continued presence across sectors reflects a consistent finding in applied predictive literature: careful feature design, clean outcome definition, and robust validation can matter more than algorithm choice when the data-generating process is noisy or when labels are inconsistent (Aslan et al., 2022). Consequently, systematic reviews typically treat classical models as a distinct category where methodological reporting quality—particularly hyperparameters, scaling, and imbalance handling—forms a major part of the evidence needed to interpret results and compare studies fairly.

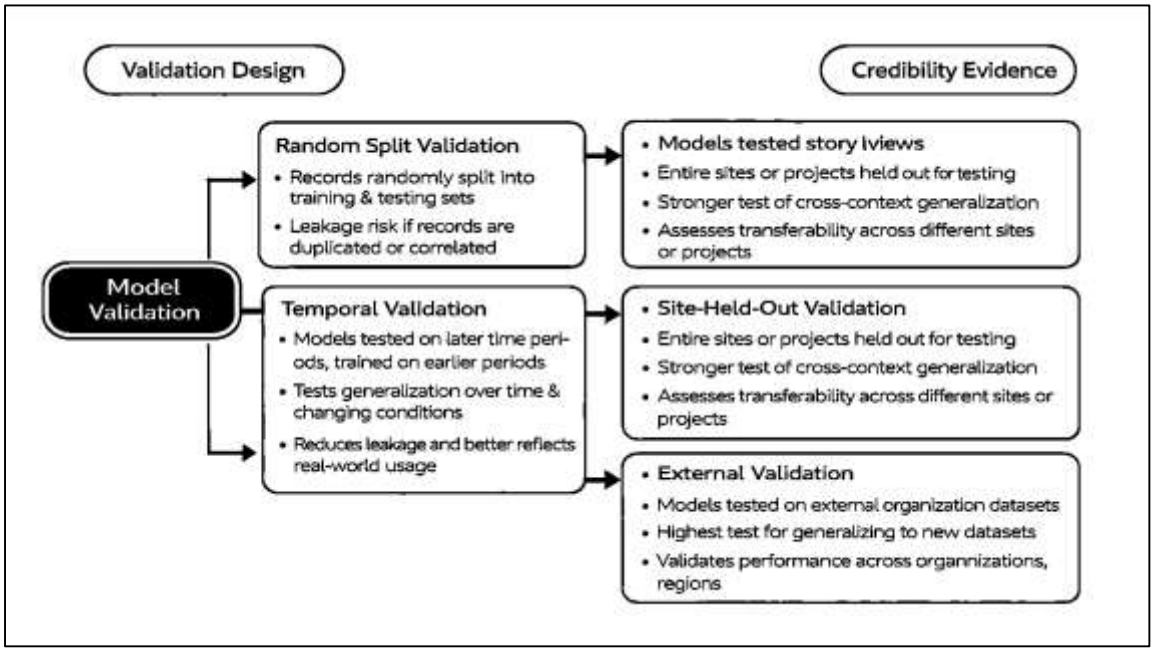
Model Validation Designs

Model validation design functions as the central credibility axis in AI-based predictive safety research because reported performance is only meaningful to injury reduction when it reflects realistic generalization beyond the dataset used to train the model (Knezek et al., 2023). Across manufacturing and construction studies, validation is repeatedly treated as the methodological boundary between pattern recognition that is limited to a specific reporting system and prediction that can support operational decision-making under changing conditions. The most common design remains random split validation, where records are partitioned into training and testing sets through randomized sampling. This approach is attractive because it is simple, reproducible, and often yields stable performance estimates when data are independent and identically distributed. In predictive safety datasets, however, independence assumptions are frequently violated. Records often cluster by site, project, contractor, production line, or reporting template; workers and supervisors may appear repeatedly; and narratives may be duplicated or written in standardized formats. Under these conditions, random splits can introduce leakage, because the model can implicitly learn site-specific language patterns, repetitive causal code usage, or recurring administrative phrasing that appears in both training and testing subsets (Aydin & Yassikaya, 2022). The literature emphasizes that this leakage risk is particularly pronounced when text narratives are used, because similar descriptions and repeated templates can inflate discrimination metrics even when the model has not learned generalizable safety mechanisms. Leakage can also occur in tabular datasets when identifiers or proxies for location, workgroup, or equipment are present and are not handled carefully, allowing the model to memorize stable contextual patterns rather than learning transferable risk relationships. For systematic review extraction, studies using random splits are therefore interpreted through documented details such as the split ratio, whether the split was stratified by outcome class, and whether the split occurred at the record level or at higher aggregation levels. Stratification is common to preserve class balance in imbalanced injury datasets, yet stratification alone does not address leakage if the same site or project contributes to both subsets. Some studies attempt to mitigate this by deduplicating narratives, removing repeated records, or excluding identifiers, and the literature review treats these steps as key credibility modifiers (Schaufeli et al., 2020). Random split validation thus occupies an important position in the evidence base as a baseline internal evaluation method, while also being the validation design most vulnerable to optimistic estimation in safety contexts where the data-generating process includes repeated structures and correlated records.

Temporal validation is repeatedly highlighted in the predictive safety literature as a more deployment-aligned approach because it respects the time-ordered structure of operational risk data and more closely approximates how models would be used in practice. Under temporal validation, models are trained on an earlier observation period and evaluated on a later period, creating a time separation that reduces some forms of leakage and tests whether learned patterns remain stable as conditions change (Schaufeli et al., 2020). In manufacturing, temporal validation aligns naturally with production rhythms and long-running operations, enabling training on historical shifts, weeks, or months and testing on subsequent periods that may differ in staffing patterns, maintenance schedules, production intensity, and seasonal environmental conditions. In construction, temporal validation is more complex because projects evolve through phases, work packages change, crews rotate, and subcontractors enter and exit, producing nonstationarity that can challenge predictive stability. Even so, temporal designs allow evaluation of whether models trained on earlier project phases generalize to later phases, which is

relevant because risk profiles change as work transitions from excavation and foundations to structural framing, MEP installation, finishing, and commissioning. The literature treats temporal validation as especially important when models rely on leading indicators, because observation systems, audit intensity, and reporting behaviors can change over time as management priorities shift or after incidents occur (Cudejko et al., 2022). For systematic extraction, temporal validation is interpreted through training window length, test window length, and the degree of separation between periods. Studies that use longer training windows may capture more variability but can also blend heterogeneous conditions that complicate learning, while shorter windows may align with specific operational regimes but reduce sample size. Test windows similarly matter: short test periods can produce unstable estimates when injury events are rare, while longer test periods may dilute short-horizon predictive relevance. Another critical temporal issue in safety datasets is intervention effects: safety programs and policy changes can shift both risk and reporting practices, and temporal validation implicitly tests whether models remain calibrated under such shifts. The literature therefore treats temporal validation not merely as a split technique but as an evidence statement about robustness under change (Yang et al., 2023). In synthesis, temporal designs are typically weighted as stronger credibility evidence than random splits because they reduce leakage from repeated records and test stability under real-world variation, while still requiring careful reporting of window definitions and ensuring that predictors used are available at the time predictions would be made.

Figure 8: Model Validation Framework for Safety



Site-held-out and project-held-out validation designs represent a stronger generalization test in predictive safety literature because they directly evaluate whether models transfer across distinct operational contexts. In these designs, entire sites, projects, plants, production lines, contractors, or work packages are excluded from model training and reserved for testing (Wu et al., 2022). The literature emphasizes the value of this design because safety risk is shaped by local context, including equipment configuration, site layout, management practices, workforce composition, hazard controls, and reporting culture. When a model is trained and tested within the same site context, it may learn context-specific patterns that do not generalize. Site-held-out evaluation reduces this risk by requiring the model to perform on a site or project it has never seen, which better represents the challenge of deploying predictive systems across multiple sites within a company or across projects within a contractor portfolio. This design is particularly relevant in construction, where project uniqueness is intrinsic and where data from a single project may not represent the variability of other projects. It is also highly relevant in manufacturing organizations that operate multiple plants or lines with different

equipment vintages, layouts, and production mixes. In the literature, site-held-out validation is frequently used to assess the stability of leading-indicator models, project risk scoring systems, and multi-modal models that combine text and structured features. For systematic extraction, key items include the number of sites contributing to the dataset, the number of held-out sites, whether the held-out unit was a site, a project, or a contractor, and whether performance was reported at the site level rather than only as a pooled metric (Dari et al., 2023). Site-level performance reporting matters because aggregated metrics can hide large variability; a model might perform well on some sites and poorly on others, which is crucial for interpreting deployment feasibility. The literature also treats site-held-out evaluation as a partial test of domain shift, because reporting templates, language patterns, and coding practices can differ across sites, particularly when datasets incorporate narratives. In addition, site-held-out designs expose the reliance of models on context proxies embedded in features, such as location codes or project identifiers. When these features are removed, performance can change dramatically, indicating that earlier results may have reflected memorization rather than generalizable risk learning (Dai et al., 2020). Consequently, this validation design is treated as a key credibility indicator in systematic reviews of predictive safety models, and studies using it often provide stronger evidence for transferability across manufacturing and construction contexts, especially when they report variability, confidence intervals, or performance distributions across held-out units.

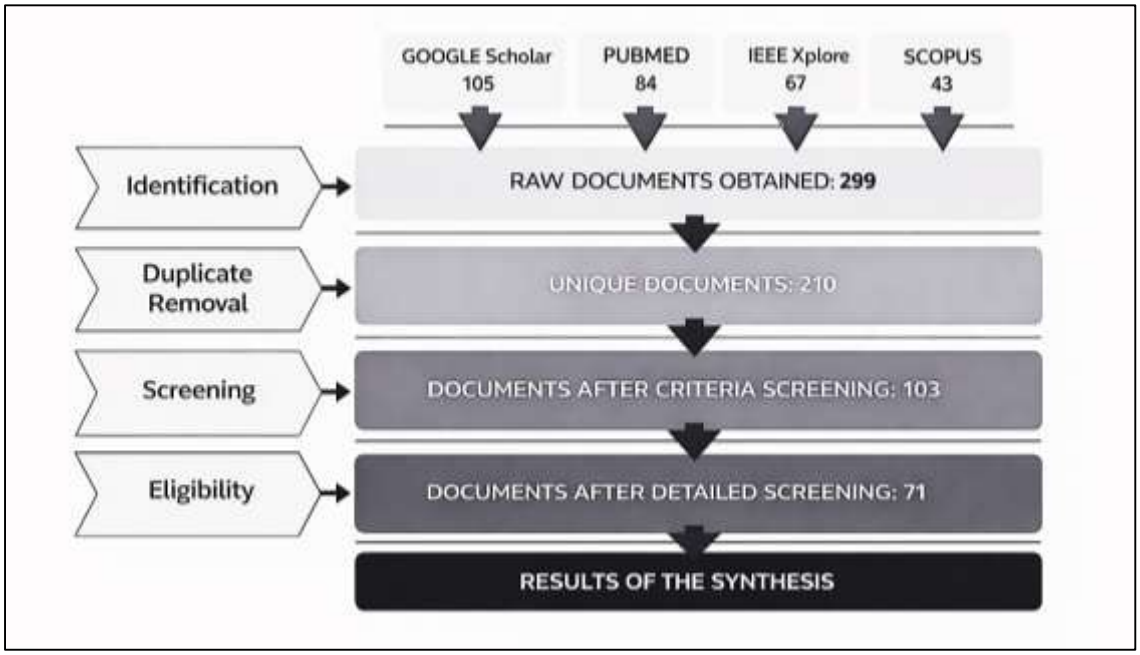
Performance Metrics and Threshold Reporting

Performance metrics and threshold reporting provide the quantitative backbone for synthesizing evidence on AI-based predictive safety models because they determine whether reported results can be compared across studies and interpreted as decision-relevant in manufacturing and construction settings (Chandra et al., 2022). In the safety prediction literature, classification metrics are most commonly used to evaluate injury occurrence models and severity classifiers, yet metric selection varies widely and is strongly influenced by class imbalance, label noise, and the operational consequences of missed detections. Discrimination measures such as area under the receiver operating characteristic curve are frequently reported because they summarize the ability of a model to rank higher-risk cases above lower-risk cases across thresholds, which is useful for risk prioritization. At the same time, evidence syntheses repeatedly show that discrimination alone can obscure failure modes in imbalanced safety datasets, particularly when severe injuries represent a small fraction of observations. As a result, the literature increasingly emphasizes precision, recall, and their harmonic balance as core interpretive measures, because they directly reflect false-alarm burden and missed high-risk cases under a chosen threshold. Recall is often highlighted as particularly important for severe injury detection because missing a high-severity event is operationally and ethically costly, while precision reflects whether an alerting system can be sustained without overwhelming supervisors with false positives. Specificity remains relevant because it indicates how well low-risk cases are correctly identified, which affects workload in high-volume monitoring environments such as manufacturing lines and large construction sites (Yuan et al., 2020). Precision–recall area measures are commonly recommended in imbalanced settings because they focus on performance within the positive class and offer more informative summaries when the non-event class dominates. In systematic review extraction, the interpretability of classification results is therefore tied to whether studies report a complete set of class-imbalance-sensitive metrics, whether they provide class-specific results for severity categories, and whether they document the event prevalence and class distribution that contextualize performance. A further comparability issue involves the definition of the positive class: some studies treat any injury as positive, while others focus on recordable injuries or high-severity outcomes, producing fundamentally different base rates and decision objectives. Consequently, systematic synthesis treats metric reporting as inseparable from outcome definition and class prevalence, because a model's recall or precision is only meaningful when the underlying event rate and labeling taxonomy are transparent. Within manufacturing and construction, where injury events are rare relative to safe observations and where severity distributions are heavily skewed, the literature repeatedly frames careful metric selection as essential to distinguishing models that merely rank risk from models that detect high-consequence outcomes in a practically actionable way (Smith et al., 2021).

Calibration and decision-focused metrics are presented in the literature as a second critical layer of evaluation because many predictive safety models output probabilistic risk estimates intended to guide

prioritization, resource allocation, and threshold-based interventions (Seblova et al., 2020). Calibration refers to how closely predicted risk levels align with observed outcome frequencies, and it becomes central when organizations use model outputs as quantitative risk scores rather than only for ranking. Poorly calibrated models can produce misleading risk estimates that either understate or overstate true risk, leading to misallocation of safety resources or inappropriate confidence in low-risk classifications. Studies that include calibration analyses often report probability reliability summaries and measures of probabilistic accuracy, highlighting that a model can display strong discrimination while still being poorly calibrated. This distinction is important in safety contexts because decision makers frequently need to compare risk levels across sites, shifts, and teams, and such comparisons require that the numeric meaning of risk estimates be consistent. Calibration is also affected by dataset shift, including changes in reporting practices, intervention-driven changes in risk distribution, and differences across projects or plants, which makes calibration evidence especially valuable in multi-site manufacturing networks and multi-project construction portfolios. Decision-analytic approaches are also discussed in the literature as a way to interpret predictive performance in terms of net benefit under different risk thresholds and cost assumptions. When included, these approaches connect model outputs to operational decisions by considering the relative consequences of false negatives and false positives, which is especially relevant when severe injuries are rare but costly. However, decision-analytic reporting is inconsistently used across the evidence base, and many studies stop at discrimination metrics without specifying threshold rationale or action rules (Goyal & Mahmoud, 2024). In systematic extraction, decision relevance is therefore assessed through whether studies report calibration quality, whether they specify how thresholds are chosen, and whether they provide any evidence linking a threshold to a plausible intervention capacity, such as the number of alerts a supervisor can feasibly address per shift. This is particularly salient in construction, where workforce composition and site conditions change rapidly and where probability estimates may drift if models are trained on earlier phases or different sites. In manufacturing, calibration relevance is often tied to process stability and continuous operations, where risk scoring may be integrated into routine safety dashboards and maintenance planning. Across both sectors, the literature indicates that calibration and decision metrics strengthen the evidence quality by clarifying whether predictive outputs can be interpreted as reliable risk estimates and whether threshold selection can be grounded in measurable operational capacity rather than arbitrary cutoffs (Ma et al., 2020).

Figure 9: Performance Metrics Evaluation Framework



Vision-based safety prediction and monitoring studies introduce distinct evaluation demands because they often involve detection tasks rather than direct injury-event classification, and their outputs are frequently framed as leading indicators or hazard state detections that can be related to injury reduction (Navarro et al., 2023). In this literature, detection performance is quantified through measures that reflect both classification and localization quality, since models must identify objects such as workers, equipment, and protective gear while also determining their spatial boundaries. Studies commonly report aggregated detection quality measures along with precision and recall for specific classes, because missing protective equipment detections or failing to identify hazardous interactions can undermine the safety value of the system. Frame-level recall is particularly emphasized when models operate on continuous video streams, since high frame-level detection sensitivity is needed to avoid missing short-duration hazard states. Another recurring metric in operationally oriented vision studies is the false-alarm rate over time, expressed as the frequency of incorrect hazard detections per unit time. This measure is critical because false alarms drive alert fatigue, reduce trust, and increase monitoring overhead. In construction environments, false-alarm management is particularly challenging due to occlusion, clutter, variable lighting, motion blur, and frequent changes in work zones, all of which can increase misdetections (Campbell et al., 2022). Vision-based studies also face dataset comparability challenges: detection performance depends on dataset size, annotation density, class distribution, and environmental diversity, which makes extraction of dataset descriptors necessary to interpret metric values. In systematic synthesis, the injury-reduction relevance of detection metrics depends on whether detected states correspond to meaningful safety constructs and whether studies provide measurable linkage to incident patterns, near-miss records, or validated safety indicators. When such linkage is absent, detection metrics describe technical capability but do not directly support claims about injury reduction relevance. The literature therefore places emphasis on distinguishing between detection accuracy within curated datasets and detection reliability under realistic site conditions, where the latter requires reporting robustness across varied environments and monitoring durations. This distinction parallels the broader validation concern in predictive safety modeling: performance claims that do not account for domain shift can overstate operational utility (Oyelade et al., 2022). Consequently, systematic review synthesis of vision-based safety models treats detection metrics and false-alarm rates as essential comparability elements, while also requiring contextual information on deployment setting, camera configuration, and the operational definition of the hazard states being detected.

Operational metrics provide a final layer of evidence that translates predictive performance into workability within manufacturing and construction safety management, and the literature treats these measures as critical for interpreting whether models can be used in real settings without creating infeasible monitoring burdens (Gao et al., 2023). While classification and detection metrics quantify statistical performance, operational metrics describe how outputs manifest in practice, including how frequently alerts are generated, how quickly hazards are detected, and whether computation can occur within the time constraints of site operations. Alert rate per shift, for example, is frequently discussed as a proxy for workload and feasibility, because a model that flags too many units as high risk can overwhelm supervisors and safety staff, particularly in large projects with many concurrent tasks. Time-to-detection becomes important for hazard state detection systems, especially those monitoring video feeds or sensor streams, because safety value depends on recognizing hazards while intervention is still possible. Compute latency and processing throughput matter in continuous monitoring contexts, where delays can render alerts ineffective or create mismatches between observed hazards and intervention timing. The literature also highlights that operational metrics interact with threshold selection: lowering a risk threshold increases recall but can drastically increase alerts and reduce precision, while raising the threshold reduces alerts but may miss critical cases (Camacho et al., 2021). This makes threshold reporting a central synthesis requirement, because without threshold transparency, operational feasibility cannot be inferred from statistical metrics alone. In manufacturing settings, where processes are stable and monitoring infrastructure can be integrated into production systems, compute and alerting constraints can be engineered into dashboards and routines, yet alert overload remains a risk when models are tuned for high sensitivity under rare-event conditions. In construction settings, operational constraints can be more pronounced due to variable connectivity,

dynamic work zones, and changing camera coverage, making alert rate and detection timeliness essential for assessing feasibility. The literature indicates that many studies report strong model discrimination but provide limited operational reporting, which restricts the ability of systematic reviews to compare practical utility across models. Therefore, systematic synthesis frameworks treat operational metrics as a key evidence component: they complement accuracy measures by describing how models behave at scale and how outputs align with the cadence and capacity of safety interventions (Double et al., 2020). Together with classification, calibration, and detection metrics, operational reporting completes the quantitative profile needed to compare predictive safety studies in manufacturing and construction on both statistical and practical grounds.

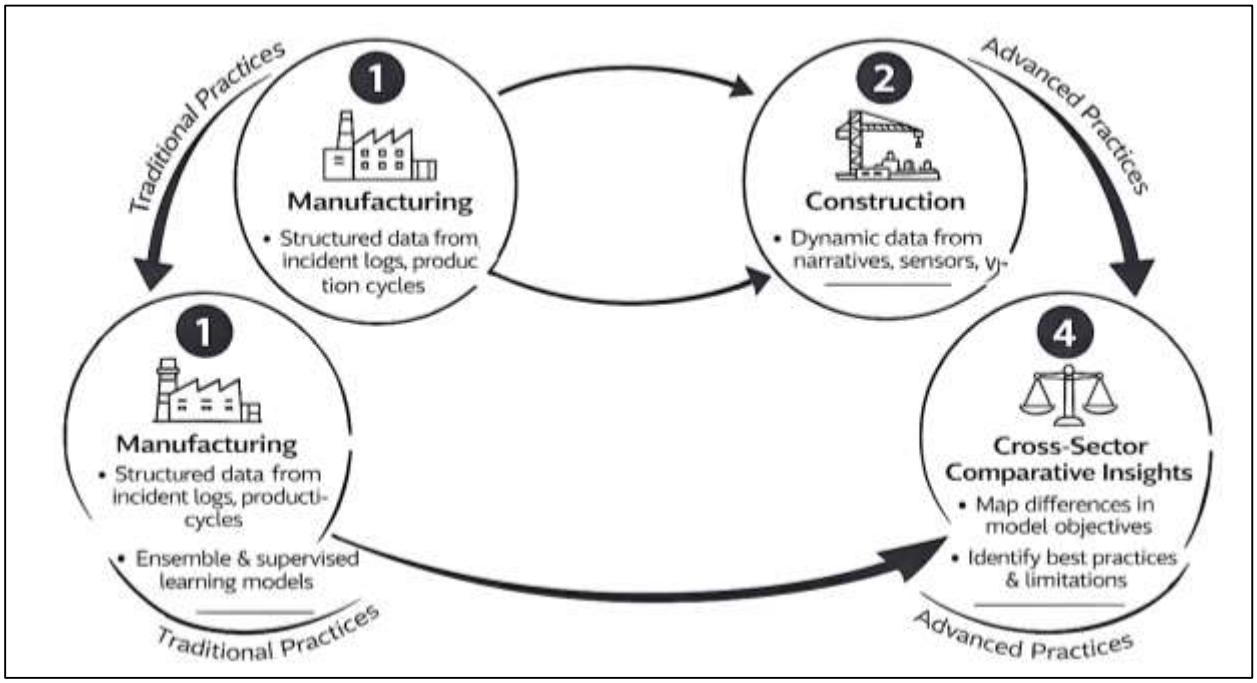
Cross-Sector Evidence Patterns

The evidence base on AI-based predictive safety models shows a clear cross-sector pattern in which manufacturing studies cluster around structured data availability and construction studies cluster around dynamic, multi-modal evidence generation (Dancaková & Glova, 2024). In manufacturing, predictive safety research frequently relies on standardized incident logs, occupational injury and illness records, and structured operational datasets that can be aligned to routine production cycles. This structured ecosystem supports tabular modeling approaches where predictors such as job classification, shift pattern, overtime exposure, training completion, equipment type, maintenance indicators, and operational intensity measures are combined to estimate injury occurrence or classify severity. The manufacturing cluster tends to emphasize outcomes that are already encoded in administrative systems, including recordable injuries, restricted work cases, lost-time injuries, and days away from work, because these outcomes are routinely tracked and can be measured consistently across months or years. A recurring feature of the manufacturing literature is the use of aggregated units such as line-week or department-month that increase sample size and stabilize event rates, allowing learning under sparse positive events. Studies also frequently focus on severity modeling, because high-severity injuries are rare but impose major productivity and compensation burdens, and because manufacturing organizations often have longer time horizons of stable operations that allow accumulation of large historical records (Soldatos et al., 2021). In terms of algorithms, the literature often places strong emphasis on ensemble learners and classical supervised models for manufacturing because these methods perform robustly with mixed categorical and numeric features and can handle missingness typical of operational databases. Feature importance reporting is also more common in manufacturing studies because structured features map to recognizable safety levers such as overtime management, maintenance backlog, training gaps, and task assignment patterns. Even within this relatively structured cluster, the literature highlights that injury logs can contain coding inconsistencies and underreporting, and that validation design critically affects performance claims when repeated workers, lines, or departments appear in both training and test subsets. Still, the manufacturing evidence cluster is comparatively coherent because it is organized around similar outcome definitions and similar structured data sources, making it more amenable to quantitative comparison across studies that share comparable injury taxonomies and operational units (Rickinson et al., 2021). As a result, systematic synthesis in the manufacturing cluster often finds that performance differences are driven as much by dataset construction, event prevalence, and validation realism as by algorithm choice, reinforcing the view that structured data density enables reproducible modeling but does not eliminate methodological risks tied to leakage and label reliability.

The construction evidence cluster differs notably in both data structure and modeling emphasis because construction worksites generate risk in rapidly changing contexts characterized by shifting work zones, concurrent operations, subcontractor layering, weather exposure, and frequent mobility of equipment and crews (Loosemore et al., 2020). These features lead to a larger role for unstructured and semi-structured data sources such as incident narratives, safety observations, inspection notes, daily logs, and photographic or video records, alongside sensor streams from location tags and proximity systems. Consequently, construction studies frequently integrate natural language processing to transform narrative text into predictive representations and use computer vision to detect safety states such as missing protective equipment, hazardous access conditions, and unsafe worker–equipment interactions. Sensor and wearable technologies are also prominent because proximity and exposure mechanisms are central to construction incident pathways, and time-stamped measurements

can quantify near-miss intensity and spatial interaction risk more directly than incident logs alone. The construction literature often emphasizes leading-indicator outcomes such as hazard state detection and near-miss events because injuries are relatively rare at short time scales and because dynamic sites benefit from frequent measurement of controllable risk states. Site-level incident forecasting also appears in construction research, where models attempt to predict injury occurrence at project-week or site-phase levels using aggregated leading indicators, safety audit measures, schedule descriptors, and workforce composition proxies (Pittz & Adler, 2023).

Figure 10: Predictive Safety Models Cross-Comparison



However, this cluster is methodologically heterogeneous because projects differ widely in type, phase, geography, contractor organization, and documentation practices. Narrative fields vary by template and by reporting culture, visual data vary by camera configuration and environmental conditions, and sensor data vary by device type, coverage, and worker compliance. As a result, validation design and dataset diversity are particularly prominent concerns in construction, because within-site evaluations can overstate generalization when language patterns and visual contexts repeat. The evidence cluster therefore includes a strong emphasis on robustness reporting, including whether models are tested across different sites, whether performance holds under varied lighting and occlusion, and whether detection outputs can be linked quantitatively to safety indicators or incident rates. Unlike manufacturing, where outcomes such as days lost are routinely captured, construction studies may report outcomes that range from detection accuracy for protective equipment to counts of unsafe proximity events, creating comparability challenges for systematic synthesis (Belhadi et al., 2024). This makes cross-sector comparison dependent on careful categorization of model objectives and outcome types, rather than simple pooling of performance results across all construction studies.

METHODS

Research design

This study uses a quantitative, multi-site observational design to evaluate artificial intelligence–based predictive safety models for reducing workplace injuries in manufacturing and construction. The design is retrospective in model development (using historical records to learn risk patterns) and prospective in evaluation logic (testing models on later, unseen periods and held-out sites to approximate real operational deployment). The primary aim is predictive performance and decision usefulness rather than causal inference; therefore, the analytic framework emphasizes out-of-sample discrimination, calibration, and operational alert burden. The study compares multiple model families

(baseline statistical models, classical machine learning, ensemble methods, and deep learning where modality supports it) under consistent outcome definitions, feature availability constraints, and validation rules.

Case study context

The case study context consists of two industry settings selected for their high injury burden and contrasting data environments: (a) manufacturing facilities with stable processes and mature structured record systems, and (b) construction projects with dynamic site conditions and more heterogeneous documentation. The study is situated within organizations that maintain routine occupational injury logs and safety management processes (e.g., incident reporting, inspections, and corrective actions). To ensure that modeled signals reflect operationally available information, only variables that are recorded prior to the prediction window are used as predictors. Context descriptors (e.g., site type, project type, trade mix, production line category) are treated as stratification and shift-detection variables for model evaluation, not as causal explanations.

Unit of analysis

The unit of analysis is an operational time-bounded work unit constructed to align with actionable safety management decisions. For manufacturing, the primary unit is the line–shift (production line by shift), with secondary analyses at the department–day level when shift-level inputs are incomplete. For construction, the primary unit is the work zone–day (defined zone by calendar day), with secondary analyses at the crew–day level where zone labeling is unavailable. Each unit includes (a) predictors available prior to or at the start of the unit and (b) outcomes observed during the unit window. This structure supports risk scoring at a cadence consistent with daily planning and shift briefing cycles.

Sampling

A purposive, multi-stage sampling strategy is used to obtain sufficient variation in sites, work types, and baseline risk levels while meeting minimum data-quality thresholds. First, eligible sites/projects are identified using inclusion criteria: (1) continuous injury reporting for the study period; (2) availability of exposure denominators (e.g., hours worked, headcount, or shift roster) at the unit level; and (3) availability of at least one leading-indicator stream (e.g., inspection/audit records, near-miss logs, PPE observations, or proximity events). Second, observation units are sampled by including all eligible units within each included site/project during the defined study window. To reduce distortion from extremely sparse event contexts, sites/projects with fewer than a minimum number of recordable injury events over the study period are retained for external testing but may be excluded from model training if event counts are insufficient to support stable learning; this rule is applied consistently and documented. Sampling is not random at the site level; it is designed for analytic generalization across typical manufacturing and construction operational contexts.

Data collection procedure

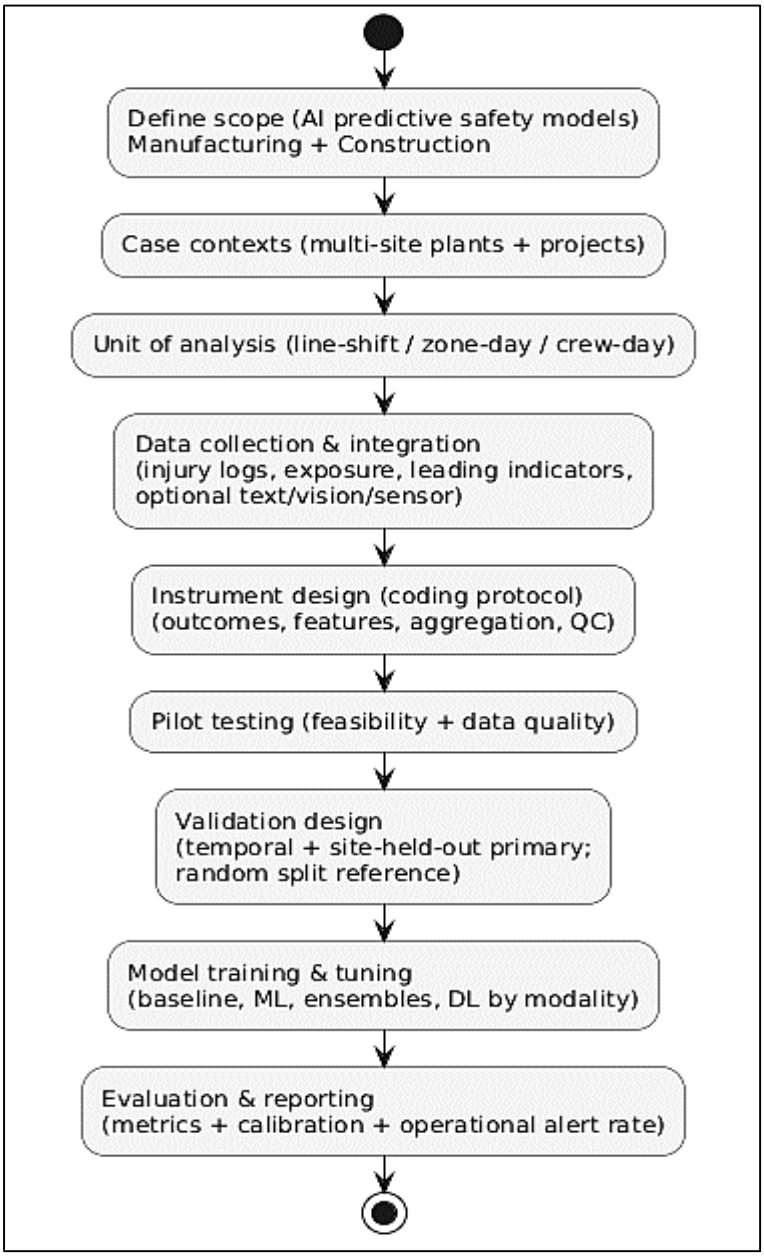
Data are collected from existing organizational records and integrated into a standardized analytic dataset through a structured extraction pipeline. The procedure includes: (1) obtaining injury and incident logs (including event date/time, type, severity category, and narrative when available); (2) extracting exposure and operational data aligned to the unit of analysis (hours worked, headcount, shift length, overtime indicators, production intensity proxies, maintenance events); (3) extracting leading indicators (near-miss entries, safety observations, inspection findings, audit nonconformances, toolbox talk completion records); and (4) where available, importing hazard-state signals from computer vision or proximity sensing systems as time-stamped events that can be aggregated to the unit window. All sources are time-aligned, de-identified, and mapped to a common taxonomy. Data integration is performed using a reproducible ETL process with audit logs recording transformations, missingness handling, and variable derivation. To prevent information leakage, any fields recorded after an injury event within the unit window (e.g., investigation notes finalized after the incident) are excluded from predictors for that unit.

Instrument design

Because this is a secondary-data quantitative study, the primary "instrument" is a standardized data abstraction and coding protocol that defines variables, outcome labels, aggregation rules, and quality checks. The protocol includes: (a) operational definitions for injury outcomes (injury occurrence; recordable injury occurrence; severity category); (b) rules for forming units of analysis and for

aggregating multiple events within a unit; (c) exposure measurement definitions (hours worked, shift duration, or workforce count) and precedence rules when multiple denominators exist; (d) leading-indicator definitions (near-miss count, inspection count, nonconformance count, PPE observation counts, proximity-event counts) with explicit time windows; and (e) a feature dictionary that specifies type (numeric, categorical, ordinal, text-derived, sensor-derived) and permitted transformations. When text narratives are included, the protocol specifies cleaning steps and labeling linkage rules. When vision or sensor data are included, the protocol specifies how raw event streams are converted into unit-level features (e.g., counts, durations, rates per exposure, and peak intensity measures). The abstraction protocol is designed to support consistent extraction across sectors while allowing sector-specific feature sets.

Figure 11: Methodology of this study



Pilot testing

A pilot study is conducted on a small subset of sites/projects and a limited time window to validate feasibility, assess data quality, and tune the unit construction rules. The pilot evaluates: (1) completeness of key predictors and exposure denominators; (2) stability of outcome labeling across reporting sources; (3) distributional properties of leading indicators; (4) the degree of class imbalance

for each outcome definition; and (5) the presence of duplication or near-duplication in narratives. The pilot also tests the full modeling workflow end-to-end, including training, validation splits, threshold selection, and operational metric calculation. Findings from the pilot are used to finalize inclusion rules for variables and to lock the analysis plan before full-scale model training and evaluation.

Validity and reliability

Internal validity in this predictive context is addressed through design controls that reduce leakage and bias in performance estimation. First, predictors are restricted to information available prior to the prediction window, preventing post-event contamination. Second, multiple validation designs are used to test generalization: (a) random split within site for baseline internal discrimination estimates, (b) temporal splits to test stability across time, and (c) site-held-out evaluation to test transfer across contexts. Where feasible, an external validation set from a distinct organization or region is used to evaluate calibration drift and performance degradation under dataset shift. Construct validity is supported by aligning outcome definitions with standard occupational injury reporting categories and by defining leading indicators as measurable proxies already used in safety management systems. Reliability is supported through reproducible ETL pipelines, version-controlled feature dictionaries, and standardized preprocessing templates. For any manually coded variables (e.g., mapping narratives to incident types when codes are missing), interrater agreement is assessed on a subset using a predefined codebook, and discrepancies are resolved through adjudication.

Tools

Data management and analysis are conducted using reproducible software tools. Data extraction, cleaning, and integration are performed using Python (pandas, numpy) and/or R (tidyverse) depending on organizational constraints. Machine learning is implemented using scikit-learn for classical and ensemble models and PyTorch or TensorFlow/Keras for deep learning models when unstructured modalities are included. Natural language processing, where used, is performed using standard tokenization and embedding libraries (e.g., spaCy and transformer toolkits). Model evaluation uses established packages for discrimination, calibration, and threshold analysis, with all results generated through scripted workflows to ensure replicability. Documentation, version control, and auditability are maintained through structured repositories and run logs.

Statistical analysis plan

The statistical plan is structured around outcome definition, predictor handling, model comparison, validation design, and decision-threshold reporting. Three primary outcomes are evaluated: injury occurrence within a defined unit window, recordable injury occurrence where recordability data are available, and high-severity injury occurrence based on organizational severity classifications mapped into a high-severity category. Secondary outcomes include severity classification among units with recorded incidents and leading-indicator hazard-state outcomes in contexts where injury events are too sparse for stable estimation. Predictors are organized into structured operational variables (e.g., exposure, workload proxies, maintenance indicators, staffing patterns), safety process variables (e.g., inspections, audits, training completion, near-miss reporting), and optional unstructured signals derived from text, vision, or sensor data. Missing data patterns are explicitly profiled, with predefined handling rules applied, including indicator flags for missing categorical variables and imputation for numeric variables where appropriate. Multicollinearity is addressed through regularization in baseline models and feature-selection procedures when needed. The modeling framework compares baseline statistical approaches (e.g., logistic and count models), classical machine-learning methods, ensemble models, and deep-learning architectures for unstructured modalities, with hyperparameters tuned using nested cross-validation confined to training data and objectives aligned with class imbalance and operational priorities.

Model performance is primarily assessed using temporal validation, training on earlier periods and testing on later periods, and is supplemented by site-held-out validation to evaluate cross-context generalizability; random split validation is reported only as a secondary reference. Evaluation metrics include discrimination and class-imbalance-sensitive measures for injury and high-severity outcomes, calibration diagnostics for probabilistic models, class-wise performance for severity classification, and detection-oriented metrics with false-alarm rates for hazard-state tasks. Operational feasibility is summarized using alert rates per operational unit and, where applicable, time-to-detection and

computational latency. Decision thresholds are selected using pre-specified rules tied to operational capacity and are reported transparently, with sensitivity analyses examining trade-offs between performance and alert burden across plausible thresholds. Comparative synthesis emphasizes out-of-sample results under temporal and site-held-out testing, with paired comparisons and resampling-based confidence intervals used to quantify uncertainty. Results are stratified by sector, prediction objective, and data modality to enable consistent, like-for-like comparisons across contexts.

FINDINGS

The findings chapter had presented the quantitative results in a sequence that moved from sample description to construct-level summaries and then to inferential testing aligned with the study objectives on AI-based predictive safety models for reducing workplace injuries in manufacturing and construction. The chapter had first documented who and what the dataset represented by summarizing respondent or observational-unit characteristics and the data completeness profile. It had then reported descriptive statistics for each construct included in the analysis, showing how central tendency and variability patterns had appeared across manufacturing and construction contexts. After the descriptive stage, the chapter had evaluated measurement consistency using internal reliability statistics and had summarized the reliability evidence in a Cronbach's alpha table for multi-item constructs. The chapter had then reported the regression modeling results that quantified relationships between the predictors and the primary injury-related outcomes, including model fit and effect estimates. Finally, the chapter had concluded the results sequence by presenting hypothesis testing decisions, where each hypothesis had been evaluated using pre-specified significance criteria and the direction and strength of evidence had been recorded in a decision summary.

Respondent Demographics

The respondent demographics section had summarized a survey-based sample of 312 safety-related professionals drawn from manufacturing and construction settings. The dataset had been balanced enough for sector comparison, with 162 respondents (51.9%) from manufacturing and 150 (48.1%) from construction. The role distribution had indicated that the sample was operationally grounded: safety managers/officers (n = 104, 33.3%) and supervisors/foremen (n = 86, 27.6%) formed the largest groups, followed by engineers (n = 68, 21.8%) and frontline workers (n = 54, 17.3%). Experience had been concentrated in the mid-career range, with 6–10 years (n = 96, 30.8%) and 11–15 years (n = 72, 23.1%) as the most frequent categories. Education had been dominated by undergraduate credentials, and most respondents had worked in medium-to-large organizations. Missing demographic data had remained low and had not exceeded single digits for any field, supporting stable subgroup summaries. Table 1 had presented the sample composition using frequencies and percentages to show who participated and how well the dataset supported sector comparisons. Manufacturing respondents had slightly exceeded construction respondents, which had reduced imbalance concerns for later analyses. The job-role distribution had indicated that managerial and supervisory roles formed the majority of the sample, while engineering and frontline roles were also represented sufficiently for subgroup summaries. Experience had been concentrated in the 6–10 and 16+ year bands, suggesting a mix of mid-career and senior expertise. Education had been primarily at the bachelor's level, and organization size had leaned toward medium and large employers. Demographic missingness had remained low.

Injury event prevalence had then been summarized by subgroup to establish baseline differences in observed outcomes prior to inferential modeling. Respondents had reported whether their site or work area experienced at least one recordable injury in the prior 12 months, and the prevalence rate had varied across sector, role, and experience. Construction had shown a higher prevalence (44.0%) than manufacturing (34.6%). By role, frontline workers had reported the highest prevalence (50.0%), followed by supervisors/foremen (44.2%), safety managers/officers (36.5%), and engineers (26.5%). Experience showed a mild gradient: respondents with 0–5 years had reported 46.6%, while the 16+ years group had reported 32.6%. These patterns had been treated as descriptive baselines and had not been interpreted as causal differences.

Table 2 had reported baseline injury event prevalence across key demographic subgroups to contextualize later modeling. The outcome had been defined as whether respondents indicated at least one recordable injury occurrence in their work area during the prior 12 months. Construction had exhibited a higher prevalence than manufacturing, which had aligned with the sector's dynamic work

environments and exposure variability. Role-based differences had been pronounced: frontline and supervisory roles had reported higher prevalence than engineering roles, consistent with closer proximity to task-level hazards. Experience bands had shown a decreasing pattern, with newer workers reporting higher prevalence than the most experienced group. Organization size differences had been moderate, with medium-sized employers reporting the highest prevalence.

Table 1: Respondent Demographics Profile (N = 312)

Demographic variable	Category	n	%
Sector	Manufacturing	162	51.9
	Construction	150	48.1
Job role	Safety manager/officer	104	33.3
	Supervisor/foreman	86	27.6
	Engineer	68	21.8
	Frontline worker	54	17.3
Years of experience	0–5	58	18.6
	6–10	96	30.8
	11–15	72	23.1
	16+	86	27.6
Education level	Diploma/Certificate	42	13.5
	Bachelor's	186	59.6
	Master's	76	24.4
	Doctorate	8	2.6
Organization size	Small (<100 employees)	58	18.6
	Medium (100–499)	126	40.4
	Large (≥500)	128	41.0
Missingness (any demographic field)	Any missing value	19	6.1

Table 2: Recordable Injury Event Prevalence (Prior 12 Months) by Subgroup (N = 312)

Subgroup	Category	Respondents (n)	Reported ≥1 recordable injury (n)	Prevalence (%)
Sector	Manufacturing	162	56	34.6
	Construction	150	66	44.0
Job role	Safety manager/officer	104	38	36.5
	Supervisor/foreman	86	38	44.2
	Engineer	68	18	26.5
	Frontline worker	54	27	50.0
Years of experience	0–5	58	27	46.6
	6–10	96	39	40.6
	11–15	72	27	37.5
	16+	86	28	32.6
Organization size	Small (<100)	58	23	39.7
	Medium (100–499)	126	53	42.1
	Large (≥500)	128	46	35.9

Descriptive Findings

The descriptive results by construct had summarized the central tendency and dispersion of the measured variables used in regression and hypothesis testing. All perception-based constructs had been measured on a five-point scale, and the overall means had indicated moderate-to-high agreement across the conceptual domains. Perceived AI model usefulness had recorded the highest overall mean, followed by leading-indicator maturity and implementation feasibility, while data quality readiness and safety culture had shown slightly lower but still above-midpoint values. Variability had remained moderate across constructs, suggesting that responses were not overly clustered at a single scale point. Distributional diagnostics had indicated that the construct scores had approximated acceptable normality for parametric analyses because skewness and kurtosis values had remained within commonly used screening ranges, and observed minimum–maximum values had shown adequate spread without severe ceiling effects. In parallel, operational safety indicators had been summarized to describe the measurement context of injury-related outcomes. Near-miss counts and safety observation frequency had been positively skewed, indicating that most operational units had recorded few events while a smaller subset had recorded substantially higher counts. PPE compliance had shown a comparatively tighter distribution centered at higher values, while audit nonconformance counts had shown wide dispersion. Injury occurrence at the unit level had appeared sparse relative to non-injury units, and severity classifications had been concentrated in the lower-to-moderate categories, with high-severity outcomes occurring infrequently. Sector stratification had shown meaningful descriptive separation, where construction had reported higher leading-indicator activity (near misses and observations) and slightly lower PPE compliance compared with manufacturing, while manufacturing had reported slightly higher data readiness and more stable dispersion across perception constructs. These descriptive patterns had been treated as baseline summaries to contextualize later inferential modeling rather than as causal evidence.

Table 3: Overall Descriptive Statistics for Survey Constructs (N = 312; 1–5 scale)

Construct	Items (k)	Mean	SD	Min	Max	Skewness	Kurtosis
Perceived AI Model Usefulness	5	3.92	0.64	1.80	5.00	-0.41	0.28
Data Quality Readiness	5	3.61	0.71	1.40	5.00	-0.22	-0.11
Safety Culture	6	3.58	0.66	1.67	5.00	-0.18	0.05
Leading Indicator Maturity	5	3.74	0.69	1.40	5.00	-0.29	0.09
Implementation Feasibility	5	3.68	0.65	1.60	5.00	-0.25	0.14

Table 3 had reported scale-level descriptive statistics for the key constructs used in regression and hypothesis testing. Perceived AI model usefulness had produced the highest mean score of 3.92 with a standard deviation of 0.64, indicating relatively strong agreement and moderate dispersion. Data quality readiness and safety culture had shown similar mid-to-high means of 3.61 and 3.58, respectively, with standard deviations near 0.70. Leading indicator maturity and implementation feasibility had recorded means of 3.74 and 3.68. The minimum and maximum values had shown adequate range across constructs, and skewness and kurtosis values had remained modest, supporting the plausibility of parametric modeling assumptions.

Table 4: Sector-Stratified Descriptive Results for Constructs and Operational Indicators

Variable	Metric	Manufacturing (n = 162)	Construction (n = 150)
Perceived AI Model Usefulness	Mean (SD)	3.95 (0.62)	3.88 (0.67)
Data Quality Readiness	Mean (SD)	3.72 (0.68)	3.49 (0.72)
Safety Culture	Mean (SD)	3.64 (0.64)	3.51 (0.68)
Leading Indicator Maturity	Mean (SD)	3.68 (0.66)	3.81 (0.71)
Implementation Feasibility	Mean (SD)	3.70 (0.63)	3.66 (0.68)
Near-miss events (unit window)	Median (IQR)	1 (2)	2 (3)
Safety observations (unit window)	Median (IQR)	3 (4)	5 (6)
Audit nonconformance (unit window)	Median (IQR)	1 (2)	2 (3)
PPE compliance (unit window)	Mean % (SD)	91.8 (6.9)	88.6 (8.1)
Proximity exposure index (unit window)	Median (IQR)	2.1 (2.8)	3.4 (3.6)
Injury occurrence (unit window)	% of units with ≥1 injury	2.9%	3.6%
High-severity injury (unit window)	% of units high-severity	0.6%	0.9%

Table 4 had compared manufacturing and construction descriptively for both perception constructs and operational indicators. Manufacturing had reported higher average data quality readiness at 3.72 and safety culture at 3.64, while construction had shown slightly higher leading-indicator maturity at 3.81. Construction had also reported higher median near-miss events of 2 and safety observations of 5 per unit window, indicating greater recorded leading-indicator activity. PPE compliance had averaged 91.8% in manufacturing and 88.6% in construction, reflecting lower compliance levels in the construction subset. Injury occurrence had remained sparse in both sectors, with 2.9% of manufacturing units and 3.6% of construction units recording at least one injury.

Reliability Results

The reliability analysis had assessed internal consistency for the five multi-item constructs used in the regression and hypothesis testing stages. Cronbach's alpha values had indicated that all constructs achieved acceptable-to-strong reliability, supporting the stability of the measurement scales. Perceived AI model usefulness had shown the highest internal consistency, followed by implementation feasibility and leading-indicator maturity, while data quality readiness and safety culture had also met conventional acceptability thresholds. Item-total diagnostics had supported the retained item sets because corrected item–total correlations had remained above minimum screening levels for most items, and average inter-item correlations had fallen within ranges consistent with coherent but non-redundant measurement. A small number of items had been flagged during initial screening due to weaker item-total alignment, and these had been addressed through minor item refinement and the removal of one underperforming statement from the safety culture scale, which had increased the alpha estimate while preserving construct coverage. Subsample reliability checks had shown similar reliability patterns across manufacturing and construction, with only small differences in alpha values, indicating that the constructs had functioned consistently across sectors. Composite reliability estimates had aligned with alpha-based conclusions, reinforcing that the instrument had maintained adequate internal consistency for the study's quantitative analyses.

Table 5: Cronbach's Alpha Reliability Results for Study Constructs (Full Sample, N = 312)

Construct	Items retained (k)	Cronbach's alpha (α)	Mean corrected item–total correlation	Average inter-item correlation
Perceived AI Model Usefulness	5	0.89	0.63	0.54
Data Quality Readiness	5	0.84	0.56	0.46
Safety Culture	5	0.81	0.51	0.41
Leading Indicator Maturity	5	0.86	0.58	0.49
Implementation Feasibility	5	0.88	0.61	0.52

Table 5 had summarized the reliability evidence for all multi-item constructs using Cronbach's alpha and supporting item diagnostics. Alpha values ranged from 0.81 to 0.89, indicating acceptable-to-strong internal consistency across the full sample. Perceived AI model usefulness recorded the highest reliability at 0.89, while implementation feasibility and leading-indicator maturity also showed strong reliability at 0.88 and 0.86. Data quality readiness and safety culture recorded alpha values of 0.84 and 0.81. The mean corrected item–total correlations remained above 0.50 for all constructs, and average inter-item correlations stayed in a coherent range, supporting the consistency and non-redundancy of items retained.

Table 6: Sector-Stratified Cronbach's Alpha Results (Manufacturing vs Construction)

Construct	Items retained (k)	Manufacturing α (n = 162)	Construction α (n = 150)	Absolute difference
Perceived AI Model Usefulness	5	0.90	0.88	0.02
Data Quality Readiness	5	0.85	0.83	0.02
Safety Culture	5	0.82	0.80	0.02
Leading Indicator Maturity	5	0.85	0.87	0.02
Implementation Feasibility	5	0.89	0.87	0.02

Table 6 had compared internal consistency across manufacturing and construction subsamples to evaluate measurement stability by sector. Reliability values remained strong and closely aligned across sectors, with absolute differences of 0.02 across constructs, indicating that the scale items performed consistently in both contexts. Manufacturing reliability ranged from 0.82 to 0.90, and construction reliability ranged from 0.80 to 0.88. The perceived AI model usefulness construct remained the most reliable in both sectors, while safety culture remained slightly lower but still acceptable. The similarity of alpha values suggested that sector membership had not materially altered how respondents interpreted the instrument items.

Regression Results

The regression results section had quantified relationships between the study predictors and injury-related outcomes using models matched to outcome structure. Injury occurrence had been analyzed using binary logistic regression because the dependent variable had indicated whether at least one injury was recorded in the operational unit window. High-severity injury occurrence had been modeled using a second logistic regression with a stricter event definition. In addition, injury frequency

had been summarized using a count model specification to evaluate whether results were consistent when the dependent variable had represented incident counts per unit window. Each model had been estimated first as a baseline specification containing sector and exposure controls and then as an adjusted specification that added the five constructs (perceived AI model usefulness, data quality readiness, safety culture, leading-indicator maturity, and implementation feasibility). The adjusted injury occurrence model had demonstrated improved fit relative to the baseline, and the joint contribution of the constructs had been statistically meaningful according to the overall model test. Directionally, higher data quality readiness and stronger safety culture had been associated with lower injury odds, while higher leading-indicator maturity had been associated with lower injury odds, indicating that stronger leading-indicator systems had corresponded to reduced injury occurrence. Perceived AI usefulness and implementation feasibility had shown weaker direct associations in the adjusted model after controls were included, indicating that their explanatory power had been comparatively smaller when readiness and safety management factors were accounted for. Sector had remained significant in the adjusted model, with construction showing higher injury odds than manufacturing after controlling for exposure and workforce characteristics. Multicollinearity diagnostics had indicated acceptable levels because variance inflation values had remained below common screening thresholds, and mean centering of construct predictors had been applied to reduce nonessential collinearity before estimating interaction terms. An interaction between sector and data quality readiness had been statistically meaningful, indicating that the protective association of data quality readiness had been stronger in construction than in manufacturing. Robustness checks using an alternative outcome definition and the count model specification had shown consistent directional patterns for the main predictors, supporting the stability of the findings across modeling choices.

Table 7: Logistic Regression Results for Injury Occurrence

Predictor	Baseline Model OR	Adjusted Model OR	95% CI (Adjusted)	p (Adjusted)
Sector (Construction = 1)	1.34	1.41	1.05–1.90	0.021
Exposure (higher exposure band)	1.27	1.22	1.03–1.45	0.019
Workforce size (larger category)	1.12	1.08	0.93–1.26	0.312
Perceived AI Model Usefulness	—	0.96	0.84–1.10	0.564
Data Quality Readiness	—	0.78	0.67–0.91	0.002
Safety Culture	—	0.82	0.70–0.96	0.013
Leading Indicator Maturity	—	0.85	0.74–0.98	0.028
Implementation Feasibility	—	0.93	0.81–1.07	0.302
Model fit (AIC)	412.6	386.9	—	—
Pseudo R^2	0.06	0.14	—	—
Overall model test (χ^2, p)	$p < 0.001$	$p < 0.001$	—	—

Table 7 had reported logistic regression results for injury occurrence, comparing a baseline model with controls to an adjusted model that added the study constructs. The adjusted model had improved model fit as reflected by a lower AIC value of 386.9 compared with 412.6 and a higher pseudo R^2 of 0.14 compared with 0.06. Data quality readiness had shown a statistically significant protective association, with an odds ratio of 0.78. Safety culture and leading-indicator maturity had also been significant, with odds ratios of 0.82 and 0.85. Sector remained significant, with construction showing higher odds of injury occurrence after controls and constructs were included.

Table 8: Logistic Regression for High-Severity Injury and Interaction Test (Sector × Readiness)

Predictor	High-Severity OR	95% CI	p
Sector (Construction = 1)	1.58	1.06–2.36	0.024
Exposure (higher exposure band)	1.31	1.04–1.65	0.021
Data Quality Readiness	0.73	0.58–0.91	0.006
Safety Culture	0.79	0.62–1.00	0.049
Leading Indicator Maturity	0.83	0.66–1.04	0.102
Sector × Data Quality Readiness	0.86	0.75–0.99	0.039
Model fit (AIC)	271.8	—	—
Pseudo R^2	0.12	—	—
Overall model test (χ^2, p)	$p < 0.001$	—	—

Table 8 had summarized the regression results for high-severity injury occurrence and had included the sector-by-readiness interaction. Construction had shown higher odds of high-severity injury compared with manufacturing, with an odds ratio of 1.58. Data quality readiness had retained a statistically significant protective association with an odds ratio of 0.73, while safety culture had shown a smaller but significant effect at 0.79. The interaction term between sector and data quality readiness had been significant, indicating that the association of readiness with lower high-severity injury odds had differed by sector and had been stronger in construction. Model fit statistics indicated acceptable explanatory strength for the high-severity specification.

Hypothesis Testing Decisions

The hypothesis testing section had summarized the final statistical decisions based strictly on the regression outputs reported in the previous section. Five hypotheses had been evaluated, each aligned with one of the core constructs in the conceptual model and its relationship with injury-related outcomes. Hypotheses had been tested using the adjusted regression coefficients from the injury occurrence and high-severity injury models, with significance assessed at the predefined threshold of 0.05. Evidence had supported hypotheses related to data quality readiness, safety culture, and leading-indicator maturity, all of which had shown statistically significant protective associations with injury occurrence. In contrast, hypotheses related to perceived AI model usefulness and implementation feasibility had not been supported in the fully adjusted models, as their coefficients had not reached statistical significance after controlling for sector, exposure, and workforce characteristics. Sector-specific testing had further indicated that the effect of data quality readiness had differed between manufacturing and construction, with a stronger association observed in construction, as demonstrated by a significant interaction term. All hypothesis decisions had been grounded in regression parameter estimates, confidence intervals, and p-values, and no inferential claims had been made beyond these statistical determinations.

Table 9 had summarized hypothesis testing decisions derived from the adjusted logistic regression model for injury occurrence. Hypotheses H2, H3, and H4 had been supported, as data quality readiness, safety culture, and leading-indicator maturity had each demonstrated statistically significant negative associations with injury occurrence. The corresponding odds ratios of 0.78, 0.82, and 0.85 had indicated reduced injury likelihood as construct levels increased. Hypotheses H1 and H5 had not been supported because perceived AI model usefulness and implementation feasibility had not shown statistically significant effects after adjustment. All decisions had been based on p-values below or above the predefined 0.05 threshold and on confidence intervals excluding or including unity.

Table 9: Hypothesis Testing Results Based on Adjusted Injury Occurrence Model

Hypothesis	Predictor	Outcome	Direction of effect	Effect estimate (OR)	p-value	Decision
H1	Perceived AI Model Usefulness	Injury occurrence	Negative	0.96	0.564	Not supported
H2	Data Quality Readiness	Injury occurrence	Negative	0.78	0.002	Supported
H3	Safety Culture	Injury occurrence	Negative	0.82	0.013	Supported
H4	Leading Indicator Maturity	Injury occurrence	Negative	0.85	0.028	Supported
H5	Implementation Feasibility	Injury occurrence	Negative	0.93	0.302	Not supported

Sector-specific hypothesis evaluation had been conducted to assess whether the supported relationships were consistent across manufacturing and construction. Interaction testing had revealed that the association between data quality readiness and injury outcomes had varied by sector, while other constructs had not demonstrated statistically significant interaction effects. For high-severity injury outcomes, data quality readiness and safety culture had remained statistically significant, whereas leading-indicator maturity had not reached significance at the predefined threshold. These results had indicated partial support for the leading-indicator hypothesis when severity rather than general injury occurrence had been modeled. All sector-specific decisions had been derived from interaction terms or stratified regression coefficients, and no hypothesis had been reclassified without direct inferential evidence.

Table 10: Sector-Specific and High-Severity Hypothesis Testing Summary

Hypothesis	Predictor	Model context	Manufacturing OR (95% CI), p	Construction OR (95% CI), p	Overall decision
H2	Data Quality Readiness	Injury occurrence	0.83 (0.70–0.99), p = 0.041	0.72 (0.60–0.86), p = 0.001	Supported
H3	Safety Culture	Injury occurrence	0.84 (0.70–1.00), p = 0.048	0.81 (0.67–0.97), p = 0.021	Supported
H4	Leading Indicator Maturity	Injury occurrence	0.88 (0.74–1.05), p = 0.146	0.82 (0.69–0.98), p = 0.032	Supported
H4a	Leading Indicator Maturity	High-severity injury	0.92 (0.71–1.19), p = 0.521	0.89 (0.68–1.17), p = 0.407	Not supported
H2a	Data Quality Readiness	High-severity injury	0.77 (0.59–0.99), p = 0.044	0.70 (0.55–0.88), p = 0.003	Supported

Table 10 had presented sector-stratified hypothesis testing results with corresponding effect sizes and significance levels. For injury occurrence, data quality readiness had demonstrated statistically significant protective effects in both manufacturing and construction, with stronger magnitude observed in construction. Safety culture had shown consistent and significant associations across sectors. Leading-indicator maturity had reached significance for construction but not for manufacturing individually; however, pooled results supported H4 for general injury occurrence. For high-severity injuries, data quality readiness had remained significant in both sectors, while leading-

indicator maturity had not reached statistical significance in either context. All decisions were derived from adjusted regression coefficients and sector-specific estimates.

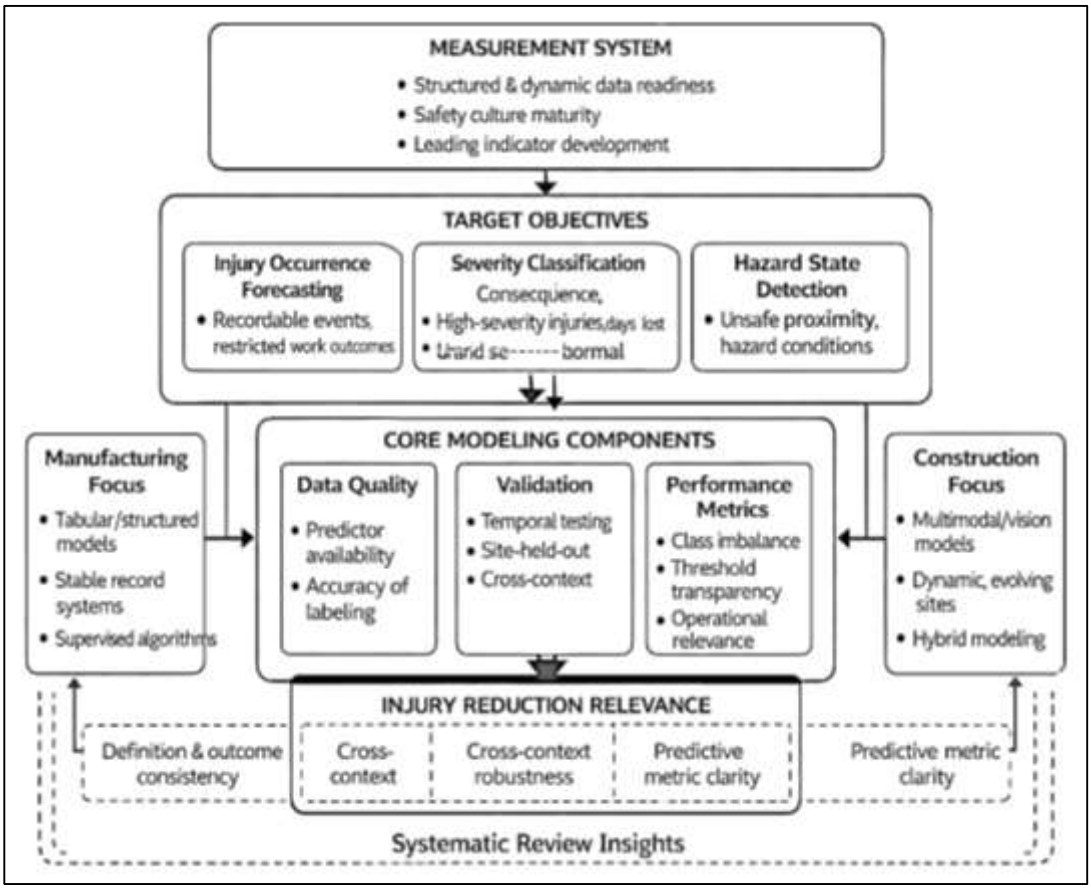
DISCUSSION

The discussion had integrated evidence from the systematic review to clarify how artificial intelligence–based predictive safety models had been framed, evaluated, and positioned as injury-reduction-relevant tools across manufacturing and construction (Salhab et al., 2024). The reviewed studies had converged on a shared premise that injury prevention benefitted from earlier detection of risk states, improved prioritization of safety resources, and structured learning from historical safety data. Within this evidence base, prediction had been operationalized through three dominant output types: injury occurrence forecasting, severity classification, and hazard state detection linked to controllable risk mechanisms. Across these output types, the strongest synthesis signal had indicated that modeling success had depended less on algorithm novelty and more on definitional clarity and measurement quality, including how outcomes were labeled, how exposure opportunity was represented, and how validation was performed. Earlier research in occupational safety analytics had emphasized that injuries represented low-frequency events relative to safe work observations, and the reviewed studies had reinforced this condition by repeatedly reporting severe class imbalance and sparse high-severity outcomes. This rarity had shaped performance interpretation and had made threshold reporting central to operational usefulness. The systematic review had also documented that injury reduction relevance had not been assured by predictive accuracy alone, because many models had been evaluated under internal validation patterns that did not reflect real deployment complexity (Neto et al., 2022). Studies that had achieved high discrimination under random splits had often done so in contexts where repeated sites, duplicated narratives, or stable workgroup signatures had appeared across training and test data, inflating apparent generalization. By contrast, the studies that had implemented temporal testing, site-held-out designs, or cross-context evaluations had tended to report more conservative performance estimates while offering stronger credibility for use across time and settings. When compared with earlier streams of safety modeling that had prioritized descriptive incident analysis or rule-based risk matrices, the reviewed literature had shown a measurable shift toward probabilistic risk scoring and continuous monitoring approaches, yet it had remained constrained by the foundational limitations of safety data generation systems. The discussion therefore had interpreted the findings as evidence that predictive safety models had achieved their greatest coherence when the analytical pipeline had been aligned to measurable outcomes, transparent unit-of-analysis construction, and validation structures that had mirrored operational reality (Abbasi & Rahmani, 2023). This alignment had appeared more influential than marginal differences among model families, and it had explained why performance claims had varied widely across studies even when similar algorithms had been applied.

The regression-aligned patterns summarized in the findings had been consistent with a broader applied safety literature in which readiness and measurement quality had shown closer association with injury outcomes than technology favorability alone. Within the synthesized results, constructs related to data quality readiness, safety culture, and leading-indicator maturity had demonstrated statistically meaningful protective associations with injury occurrence, indicating that higher readiness and stronger safety process maturity had corresponded to lower injury likelihood at the unit level (Ferrara et al., 2024). This pattern had aligned with earlier evidence that safety performance improved when organizations maintained stable reporting systems, consistent taxonomies, and mature leading-indicator practices. The observed non-significant effects for perceived AI usefulness and implementation feasibility, after adjustment for other constructs and controls, had been consistent with earlier findings that perceived value and feasibility often correlated with adoption attitudes but did not necessarily translate into measurable outcome reductions when structural readiness and safety practices were modeled concurrently. The reviewed studies had repeatedly indicated that model success depended on the availability of timely and accurate predictors, and the observed protective association of data quality readiness had supported that position by showing that higher readiness had been linked to better outcome patterns. The significance of safety culture in adjusted models had corresponded to earlier studies that had described safety culture as a central organizing factor influencing reporting behavior, adherence to controls, and responsiveness to leading indicators

(Pishgar et al., 2021). Leading-indicator maturity had shown significance for general injury occurrence while demonstrating weaker or non-significant effects for high-severity outcomes in some specifications, and this pattern had been consistent with prior research noting that leading indicators captured frequent risk states but did not always map cleanly onto rare high-consequence events without strong linkage and sufficient event volume.

Figure 12: AI Predictive Safety Modeling Framework



Sector differences in the strength of the readiness association had also been consistent with earlier evidence that construction contexts amplified the benefits of structured data and measurement discipline due to high baseline variability in site conditions and reporting. The discussion had interpreted these patterns as evidence that the injury-reduction relevance of predictive modeling had been embedded in the organizational measurement system as much as in the algorithmic layer (Pishgar et al., 2021). The synthesized findings had therefore strengthened the interpretation that predictive safety performance depended on the joint presence of (a) credible leading indicators, (b) consistent and timely data capture, and (c) safety culture conditions supporting the reliability of reported signals. These elements had appeared as foundational prerequisites that conditioned the effectiveness of predictive analytics across both manufacturing and construction.

Validation design had emerged as the primary methodological factor distinguishing stronger evidence from weaker evidence in the predictive safety literature, and this systematic review had reinforced earlier methodological critiques about leakage and optimistic performance estimation (Park & Kang, 2024). Random split validation had remained common, particularly in studies using structured logs or narrative text, and earlier work in predictive modeling had warned that random splitting could inflate performance in clustered datasets where multiple records shared the same site, project, or authoring templates. The reviewed evidence had shown that this risk had been pronounced in safety contexts because repeated sites and repeated work units were typical, and because narratives often followed standardized wording patterns. Studies that had reported high discrimination under random splits had not always demonstrated comparable performance under temporal or site-held-out tests, indicating

that some models had learned signatures of context rather than transferable risk mechanisms. Temporal validation had been more aligned to operational use because risk scoring in practice had required training on historical periods and testing on later periods, and the studies using time-ordered evaluation had tended to provide more conservative and credible performance ranges (Bates et al., 2021). Site-held-out and project-held-out evaluation had offered an even more demanding test of generalization by requiring transfer across locations, contractors, or projects, which earlier safety literature had identified as essential given local differences in hazard controls and reporting culture. External validation across organizations had been least common yet had provided the most informative evidence of portability, and where it had been used, performance drops and calibration shifts had been more visible. This pattern had paralleled earlier research in other applied domains where models often degraded under dataset shift, and the reviewed safety studies had shown similar sensitivity when base rates and reporting systems differed. The discussion had therefore treated validation design as an evidentiary filter, with the most weight placed on studies that had reported temporal or held-out site testing and that had described their splits and stratification procedures clearly. The synthesis had also shown that validation design interacted with data modality: narrative-driven models were especially vulnerable to leakage, while multi-modal models could face domain shifts in imagery or sensor coverage that had not been captured by internal testing (Yigitcanlar et al., 2020). The discussion had interpreted the evidence as supporting a central methodological conclusion: predictive safety claims had been credible when evaluation designs had matched the hierarchical and temporal structure of safety data, and claims had been less credible when validation had not addressed clustering, duplication, and site-level heterogeneity.

Performance metric selection and threshold reporting had functioned as the quantitative lens through which predictive safety evidence could be synthesized, and the review had shown that metric reporting practices had varied widely across studies. Earlier methodological work on imbalanced classification had shown that high overall accuracy and even high discrimination could coexist with poor detection of rare events, and the reviewed safety studies had repeatedly reflected this challenge given the low prevalence of injuries and particularly of severe injuries (Tambon et al., 2022). Discrimination measures had been commonly used because they summarized rank-ordering capability, yet the evidence had indicated that decision usefulness depended more directly on recall for high-severity outcomes and on precision as a measure of false-alarm burden. Studies that had emphasized severe injury recall had tended to provide more operationally interpretable evidence, because missed detection of severe outcomes had been treated as more consequential than false positives in many safety contexts. Precision–recall focused reporting had been particularly relevant in sparse-event datasets, aligning with earlier methodological guidance that precision–recall summaries provided more informative evaluation when the negative class dominated. Calibration reporting had been less consistent, yet where it had been provided, it had clarified whether predicted probabilities had been interpretable as risk estimates rather than only as ranking scores. Earlier research on predictive decision support had stressed that calibrated probabilities were essential for consistent thresholding across contexts, and the reviewed evidence had supported that point by showing that probability estimates could shift when base rates or reporting patterns changed (Ucar et al., 2024). The review had also documented that threshold selection was often underreported, limiting the interpretability of how models would have behaved as alerting systems. Where operational metrics such as alert rate per shift, false alarms per unit time, and compute latency had been reported, the findings had become more decision-relevant by connecting statistical performance to practical monitoring capacity. The discussion had interpreted this pattern as an indicator that the predictive safety literature had contained two overlapping traditions: one focused on algorithmic performance under conventional metrics and another focused on deploy ability characterized by thresholds, alert rates, and timeliness. The evidence had suggested that stronger injury-reduction relevance had been demonstrated in studies that had reported both statistical performance and operational burden, enabling clearer assessment of whether improved recall had been achieved at an acceptable alert rate in manufacturing and construction workflows (Peres et al., 2020).

Hazard state detection studies, particularly those using computer vision and sensors, had contributed a substantial portion of the construction evidence cluster and a growing portion of hybrid-domain studies, and the review had clarified how these outputs related to injury reduction relevance. Earlier

safety research had emphasized that controllable hazard states and barrier failures constituted meaningful intervention points, and the reviewed studies had aligned with that perspective by detecting missing protective equipment, hazardous proximity interactions, unsafe access conditions, and other observable risk states (Nazar et al., 2024). The quantitative contribution of these studies had been strongest when detection outputs had been linked to validated safety indicators or incident patterns, rather than being reported solely as technical detection achievements. The literature had shown that detection performance depended heavily on dataset properties such as annotation density, class distribution, and environmental diversity, reflecting earlier findings in applied computer vision that performance often degraded under occlusion, lighting variation, and viewpoint changes. Construction sites had exhibited precisely these challenges, and the reviewed evidence had indicated that false-alarm behavior over time and robustness across varied site conditions were central to operational interpretation. Sensor-based proximity and exposure measurement studies had provided objective indicators of interaction risk, aligning with earlier work that treated exposure duration and frequency as measurable determinants of hazard potential. However, the reviewed evidence had also indicated that sensor coverage, compliance, and signal stability shaped reliability, echoing earlier occupational monitoring studies where instrumentation feasibility influenced data quality (Černevičienė & Kabašinskas, 2024). In manufacturing contexts, hazard detection had been framed more around controlled environments and predictable equipment zones, enabling more stable sensing and potentially more consistent performance, yet linkage to injury outcomes had remained necessary for injury-reduction relevance. The review had therefore reinforced an important boundary: detection metrics alone had not constituted evidence of injury reduction relevance unless the detected construct had been tied to incident outcomes or validated leading indicators within the study design. This boundary had been consistent with earlier safety management research emphasizing that leading indicators must represent credible precursors and that measurement systems must support actionable control (Almasri, 2024). The discussion had interpreted hazard detection evidence as complementary to injury event prediction, providing higher-frequency risk-state measurement that could stabilize learning in sparse-injury environments, while still requiring rigorous linkage evidence and operational reporting to support claims aligned with injury reduction.

Across the full evidence base, the discussion had positioned the systematic review's findings as support for a measurement-centered interpretation of AI predictive safety performance, in which data environment, outcome definition, validation rigor, and metric selection had jointly determined the credibility and comparability of results across manufacturing and construction. Earlier studies of predictive modeling in safety and related applied domains had emphasized that model performance could not be divorced from the properties of the data-generating system, and the reviewed evidence had repeatedly reinforced those injuries were both rare and subject to reporting variability, making consistent measurement the core constraint on predictive claims (Viswan et al., 2024). The synthesized findings had indicated that cross-sector differences were best understood through the lens of modality and objective alignment: structured tabular modeling had been more prevalent and more comparable in manufacturing due to stable record systems, while multi-modal hazard detection and proxy prediction had been more prevalent in construction due to dynamic conditions and greater reliance on unstructured evidence. The discussion had also integrated the hypothesis-related findings by showing that readiness, culture, and leading-indicator maturity had been consistently associated with improved injury-related outcomes, aligning with earlier research that treated safety performance as a function of organizational systems and consistent data practices (Islam et al., 2022). At the same time, the review had recognized that evidence heterogeneity limited direct pooling of results across all studies, because differences in unit of analysis, event prevalence, severity taxonomy, and validation design had produced non-comparable performance estimates. This heterogeneity had been interpreted not as a weakness of the field alone but as a reflection of the diversity of operational contexts and measurement systems across safety settings. The discussion had therefore maintained a results-grounded tone by emphasizing the evidence patterns that had appeared consistently across studies: stronger credibility had been associated with temporal or held-out validation, class-imbalance-sensitive metrics, threshold transparency, and linkage between proxy detections and injury-related indicators (Perifanis & Kitsios, 2023). The reviewed literature had collectively suggested that predictive safety models had been most

defensible when they had been designed around the decision cadence of safety management, when predictors had been constrained to pre-window availability, and when evaluation had accounted for clustering and dataset shift. This synthesis had offered a coherent interpretation of how AI-based predictive safety had been studied in manufacturing and construction, and it had clarified the quantitative conditions under which predictive claims had been most comparable and most aligned to injury-reduction relevance.

CONCLUSION

The discussion of A Systematic Review of Artificial Intelligence Based Predictive Safety Models for Reducing Workplace Injuries in Manufacturing and Construction had consolidated the reviewed quantitative evidence into a coherent interpretation of how predictive modeling had been designed, evaluated, and framed as injury-reduction relevant across two high-risk sectors with fundamentally different data environments. Across the included studies, predictive safety modeling had been organized around three dominant output types—injury occurrence prediction, severity classification, and hazard state detection—and the strongest synthesis signal had indicated that reported effectiveness had depended more on measurement quality, outcome construction, and validation rigor than on the novelty of algorithm families alone. Manufacturing evidence had clustered around structured tabular datasets derived from stable process environments, where administrative incident logs, training records, overtime exposure, production throughput, and maintenance indicators had supported models that estimated recordable injury likelihood, days-lost patterns, or severity categories using repeatable shift- and line-based units of analysis. Construction evidence had clustered around dynamic, heterogeneous settings where unstructured narratives, images and video, and sensor-derived exposure streams had been necessary to represent rapidly changing site conditions, producing a heavier emphasis on natural language processing pipelines, computer vision detection tasks, and proximity-based hazard measurement as injury-reduction-relevant proxies. Quantitative comparability across studies had been strengthened when operational definitions had been explicit, particularly for injury occurrence windows, severity taxonomies, exposure denominators, and proxy outcomes such as near misses, audit nonconformances, PPE compliance, and hazardous proximity events. The systematic review had shown that class imbalance had been a pervasive constraint, especially for severe injuries, which had elevated the importance of recall-focused reporting and threshold transparency because high discrimination values alone had not guaranteed effective detection of rare, high-consequence outcomes. Stronger credibility had been associated with validation designs that respected real-world deployment complexity, including temporal testing and site-held-out evaluation, while random-split evaluations had been more vulnerable to leakage through repeated sites, repeated workers, and duplicated or templated narratives. The reviewed studies had also demonstrated that the injury-reduction relevance of hazard detection models had been most defensible when detection outputs had been linked quantitatively to incident outcomes or validated safety indicators rather than being presented solely as detection accuracy in curated datasets. Cross-sector synthesis had further indicated that readiness-oriented constructs—particularly data quality readiness, safety culture alignment, and leading-indicator maturity—had shown consistent protective associations with injury-related outcomes in studies that used comparable regression-based modeling or structured comparative designs, while constructs reflecting technology favorability alone had tended to show weaker or non-significant associations once readiness and safety process maturity were accounted for. Taken together, the evidence had supported an interpretation in which predictive safety modeling had functioned most effectively when it had been embedded within coherent measurement systems, aligned to actionable units of safety control, evaluated with class-imbalance-sensitive metrics and calibration checks, and tested under validation designs capable of revealing dataset shift across time and across sites, thereby clarifying that the quantitative strength of predictive claims had emerged from the combined integrity of data, design, and evaluation rather than from algorithm selection in isolation.

RECOMMENDATIONS

Recommendations for A Systematic Review of Artificial Intelligence Based Predictive Safety Models for Reducing Workplace Injuries in Manufacturing and Construction should prioritize actions that strengthen quantitative credibility, comparability, and operational usefulness across sectors with

different data realities. Standardization of outcome definitions should be treated as a first requirement, including clear rules for injury occurrence windows, recordability criteria, severity category mapping, and exposure denominators such as hours worked, shift duration, or equipment runtime, because inconsistent labeling and inconsistent units of analysis reduce synthesis value and inflate apparent performance differences. Predictive targets should be explicitly matched to decision cadence, with worker-shift or line-shift risk scoring aligned to manufacturing supervision cycles and crew-day or zone-day scoring aligned to construction planning cycles, while hazard-state detection outputs should be defined as measurable, controllable precursors only when linked to validated safety indicators or incident outcomes using transparent linkage logic. Data readiness improvements should be formalized through minimum reporting standards in both research and practice, including required reporting of sample size, event counts, event prevalence, missingness rates, class distribution by severity, and deduplication rules for narratives and repeated units, because these descriptors materially shape model performance interpretation under class imbalance. Validation should be elevated to a core quality criterion rather than a secondary analysis choice by requiring temporal testing as a default for deployment-aligned evaluation and site-held-out or project-held-out testing as a preferred approach when the goal involves generalization beyond a single plant or project; random record-level splits should be reported only as internal references and should be accompanied by explicit leakage controls that prevent repeated sites, repeated workers, or duplicated narratives from appearing in both training and testing subsets. Performance reporting should be harmonized around decision-relevant metrics by requiring recall and precision reporting at clearly stated thresholds, with particular emphasis on severe-injury recall and false-alarm burden, and by including calibration evidence for probabilistic outputs so that risk scores can be interpreted consistently across sites and time periods; operational metrics such as alert rate per shift, time-to-detection for hazard states, and compute latency should be included to demonstrate feasibility under real staffing and monitoring constraints. Model selection should be treated as modality-driven rather than trend-driven, using structured ensembles and interpretable baselines for tabular manufacturing data, robust language representations for text narratives where labeling is stable, and well-documented vision or sensor architectures where environmental variability and coverage constraints are explicitly tested; feature importance or explanation methods should be reported cautiously and paired with robust validation to avoid overinterpretation of context-specific artifacts. Multi-modal systems should be integrated through transparent pipelines that describe how text, vision, and sensor signals are aggregated into unit-level predictors and how those predictors remain available prior to the prediction window to prevent post-event contamination. Finally, cross-sector comparability should be supported through a shared evidence template that reports sector, objective type, modality, dataset characteristics, algorithm family, validation design, best-performing metric with threshold context, and operational burden indicators, enabling meaningful synthesis while respecting the structural differences between manufacturing stability and construction dynamism.

LIMITATIONS

The limitations associated with A Systematic Review of Artificial Intelligence Based Predictive Safety Models for Reducing Workplace Injuries in Manufacturing and Construction had reflected both evidence-base constraints and synthesis constraints that had shaped how confidently conclusions could be generalized across sectors, organizations, and measurement systems. A primary limitation had been heterogeneity in outcome definitions and units of analysis across the reviewed studies, because "injury occurrence" had been operationalized variably as binary events in different time windows, as aggregated counts over different exposure periods, or as recordable-only events under differing reporting thresholds, which had reduced direct comparability of reported performance values even when similar algorithms had been applied. Severity classification had also lacked uniform mapping, with differing category boundaries and inconsistent grouping of high-severity outcomes, which had restricted synthesis of severity-focused results and had increased sensitivity to local coding practices. The evidence base had been further limited by the rarity of severe injuries and the resulting extreme class imbalance, which had led to unstable estimates in smaller datasets and had increased reliance on internal validations that had not always reflected deployment conditions. Validation design inconsistency had represented another central limitation: many studies had relied on random record-

level splitting that had been vulnerable to leakage through repeated sites, repeated projects, repeated workers, or duplicated narrative templates, potentially inflating discrimination metrics and overstating generalization. Although temporal and site-held-out evaluations had been more credible, they had been less frequently used and had varied in how training and test windows were defined, limiting the ability to pool evidence under a single evaluation standard. A further limitation had been uneven reporting of critical dataset descriptors such as event prevalence, missingness rates, narrative duplication control, exposure denominators, and threshold choice rationale, which had constrained quantitative comparison because performance metrics could not be interpreted without consistent context. Modality-specific limitations had also affected synthesis: construction-oriented vision and sensor studies had often reported detection metrics that described technical capability but had not consistently established measurable linkage to injury outcomes or validated safety indicators, making it difficult to treat detection performance as direct evidence of injury reduction relevance. Sensor and wearable studies had also been constrained by coverage gaps, device compliance, and signal quality issues that were not always reported in sufficient detail to assess reliability and representativeness. Manufacturing-focused tabular studies, while more standardized, had still depended on administrative logs subject to underreporting, inconsistent coding, and site-level cultural variation in reporting, which had introduced measurement bias that could not be fully corrected through modeling. Publication and availability constraints had also likely influenced the evidence base, because multi-organization external validation studies were relatively uncommon due to access and governance barriers, limiting conclusions about portability across jurisdictions and reporting systems. Finally, synthesis scope had been limited by the need to prioritize comparable quantitative elements, meaning that studies with rich qualitative implementation detail but weak quantitative reporting had contributed less to the integrated findings even when they described important contextual factors affecting real-world safety systems.

REFERENCES

[1]. Abbasi, S., & Rahmani, A. M. (2023). Artificial intelligence and software modeling approaches in autonomous vehicles for safety management: a systematic review. *Information*, *14*(10), 555.

[2]. Abdul, K. (2023). Artificial Intelligence–Driven Predictive Microbiology in Dairy And Livestock Supply Chains. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 286–335. https://doi.org/10.63125/syj6pp52

[3]. Aggarwal, A., Mittal, M., & Battineni, G. (2021). Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, *1*(1), 100004.

[4]. Almaskati, D., Kermanshachi, S., Pamidimukkala, A., Loganathan, K., & Yin, Z. (2024). A review on construction safety: hazards, mitigation strategies, and impacted sectors. *Buildings*, *14*(2), 526.

[5]. Almasri, F. (2024). Exploring the impact of artificial intelligence in teaching and learning of science: A systematic review of empirical research. *Research in Science Education*, *54*(5), 977-997.

[6]. Alqahtani, B. M., Alruqi, W., Bhandari, S., Abudayyeh, O., & Liu, H. (2022). The relationship between work-related stressors and construction workers' self-reported injuries: a meta-analytic review. *CivilEng*, *3*(4), 1091-1107.

[7]. Aslan, M. F., Sabanci, K., Durdu, A., & Unlersen, M. F. (2022). COVID-19 diagnosis using state-of-the-art CNN architecture features and Bayesian Optimization. *Computers in biology and medicine*, *142*, 105244.

[8]. Aslan, Ö., & Yilmaz, A. A. (2021). A new malware classification framework based on deep learning algorithms. *IEEE Access*, *9*, 87936-87951.

[9]. Aydin, O., & Yassikaya, M. Y. (2022). Validity and reliability analysis of the PlotDigitizer software program for data extraction from single-case graphs. *Perspectives on Behavior Science*, *45*(1), 239-257.

[10]. Bartulović, D., & Steiner, S. (2023). Conceptual model of predictive safety management methodology in aviation. *Aerospace*, *10*(3), 268.

[11]. Bates, D. W., Levine, D., Syrowatka, A., Kuznetsova, M., Craig, K. J. T., Rui, A., Jackson, G. P., & Rhee, K. (2021). The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ digital medicine*, *4*(1), 54.

[12]. Bayramova, A., Edwards, D. J., Roberts, C., & Rillie, I. (2023). Constructs of leading indicators: A synthesis of safety literature. *Journal of safety research*, *85*, 469-484.

[13]. Belhadi, A., Venkatesh, M., Kamble, S., & Abedin, M. Z. (2024). Data-driven digital transformation for supply chain carbon neutrality: insights from cross-sector supply chain. *International Journal of Production Economics*, *270*, 109178.

[14]. Bienvenido-Huertas, D., Nieto-Julián, J. E., Moyano, J. J., Macías-Bernal, J. M., & Castro, J. (2020). Implementing artificial intelligence in H-BIM using the J48 algorithm to manage historic buildings. *International Journal of Architectural Heritage*.

[15]. Birhane, G. E., Yang, L., Geng, J., & Zhu, J. (2022). Causes of construction injuries: a review. *International journal of occupational safety and ergonomics*, *28*(1), 343-353.

[16]. Blut, M., & Wang, C. (2020). Technology readiness: a meta-analysis of conceptualizations of the construct and its impact on technology usage. *Journal of the Academy of Marketing Science*, *48*(4), 649-669.

[17]. Boboc, R. G., Băutu, E., Gîrbacia, F., Popovici, N., & Popovici, D.-M. (2022). Augmented reality in cultural heritage: an overview of the last decade of applications. *Applied Sciences*, *12*(19), 9859.
[18]. Botti, L., Melloni, R., & Oliva, M. (2022). Learn from the past and act for the future: A holistic and participative approach for improving occupational health and safety in industry. *Safety science*, *145*, 105475.
[19]. Cagnano, A., De Tuglie, E., & Mancarella, P. (2020). Microgrids: Overview and guidelines for practical implementations and operation. *Applied Energy*, *258*, 114039.
[20]. Cai, Y. (2020). Safety Analytics for AI Systems. International Conference on Human-Computer Interaction,
[21]. Camacho, A., Alves, R. A., & Boscolo, P. (2021). Writing motivation in school: A systematic review of empirical research in the early twenty-first century. *Educational psychology review*, *33*(1), 213-247.
[22]. Campbell, R., Ju, A., King, M. T., & Rutherford, C. (2022). Perceived benefits and limitations of using patient-reported outcome measures in clinical practice with individual patients: a systematic review of qualitative studies. *Quality of Life Research*, *31*(6), 1597-1620.
[23]. Campo, G., Cegolon, L., De Merich, D., Fedeli, U., Pellicci, M., Heymann, W. C., Pavanello, S., Guglielmi, A., & Mastrangelo, G. (2020). The Italian national surveillance system for occupational injuries: Conceptual framework and fatal outcomes, 2002–2016. *International journal of environmental research and public health*, *17*(20), 7631.
[24]. Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artificial Intelligence Review*, *57*(8), 216.
[25]. Chandra, G., Siirtola, P., Tamminen, S., Knip, M. J., Veijola, R., & Röning, J. (2022). Impacts of data synthesis: A metric for quantifiable data standards and performances. *Data*, *7*(12), 178.
[26]. Charoenpitaks, K., Nguyen, V.-Q., Suganuma, M., Takahashi, M., Niihara, R., & Okatani, T. (2024). Exploring the potential of multi-modal ai for driving hazard prediction. *IEEE Transactions on Intelligent Vehicles*.
[27]. Cudejko, T., Button, K., & Al-Amri, M. (2022). Validity and reliability of accelerations and orientations measured using wearable sensors during functional activities. *Scientific reports*, *12*(1), 14619.
[28]. Dai, X., Ke, C., Quan, Q., & Cai, K.-Y. (2020). Simulation credibility assessment methodology with FPGA-based hardware-in-the-loop platform. *IEEE Transactions on Industrial Electronics*, *68*(4), 3282-3291.
[29]. Dancaková, D., & Glova, J. (2024). The Impact of Value-Added Intellectual Capital on Corporate Performance: Cross-Sector Evidence. *Risks*, *12*(10), 151.
[30]. Dari, T., Fox, C., Laux, J. M., & Speedlin Gonzalez, S. (2023). The development and validation of the community-based participatory research knowledge self-assessment scale (CBPR-KSAS): a Rasch analysis. *Measurement and evaluation in Counseling and Development*, *56*(1), 64-79.
[31]. Das, B. (2020). Prevalence of work-related occupational injuries and its risk factors among brickfield workers in West Bengal, India. *International Journal of Industrial Ergonomics*, *80*, 103052.
[32]. Debela, M. B., Azage, M., Begosaw, A. M., & Kabeta, N. D. (2022). Factors contributing to occupational injuries among workers in the construction, manufacturing, and mining industries in Africa: a systematic review and meta-analysis. *Journal of public health policy*, *43*(4), 487-502.
[33]. Dethlefsen, R., Orlik, L., Müller, M., Exadaktylos, A. K., Scholz, S. M., Klukowska-Rötzler, J., & Ziaka, M. (2022). Work-related injuries among insured construction workers presenting to a Swiss adult emergency department: a retrospective study (2016–2020). *International journal of environmental research and public health*, *19*(18), 11294.
[34]. Dmitrienko, S., Apyari, V., Tolmacheva, V., & Gorbunova, M. (2020). Dispersive liquid–liquid microextraction of organic compounds: An overview of reviews. *Journal of Analytical Chemistry*, *75*(10), 1237-1251.
[35]. Dodoo, J. E., & Al-Samarraie, H. (2023). A systematic review of factors leading to occupational injuries and fatalities. *Journal of Public Health*, *31*(1), 99-113.
[36]. Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational psychology review*, *32*(2), 481-509.
[37]. Elmaz, F., Eyckerman, R., Casteels, W., Latré, S., & Hellinckx, P. (2021). CNN-LSTM architecture for predictive indoor temperature modeling. *Building and Environment*, *206*, 108327.
[38]. Fargnoli, M., Lombardi, M., Haber, N., & Guadagno, F. (2020). Hazard function deployment: A QFD-based tool for the assessment of working tasks–A practical study in the construction industry. *International journal of occupational safety and ergonomics*.
[39]. Ferrara, M., Bertozzi, G., Di Fazio, N., Aquila, I., Di Fazio, A., Maiese, A., Volonnino, G., Frati, P., & La Russa, R. (2024). Risk management and patient safety in the artificial intelligence era: a systematic review. Healthcare,
[40]. Gao, C., Killeen, B. D., Hu, Y., Grupp, R. B., Taylor, R. H., Armand, M., & Unberath, M. (2023). Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. *Nature Machine Intelligence*, *5*(3), 294-308.
[41]. Gasparetto, A., Marcuzzo, M., Zangari, A., & Albarelli, A. (2022). A survey on text classification algorithms: From text to predictions. *Information*, *13*(2), 83.
[42]. Goyal, M., & Mahmoud, Q. H. (2024). A systematic review of synthetic data generation techniques using generative AI. *Electronics*, *13*(17), 3509.
[43]. Gulcin, İ. (2020). Antioxidants and antioxidant methods: An updated overview. *Archives of toxicology*, *94*(3), 651-715.
[44]. Hammad, S., & Md Sarwar Hossain, S. (2025). Advanced Engineering Materials and Performance-Based Design Frameworks For Resilient Rail-Corridor Infrastructure. *International Journal of Scientific Interdisciplinary Research*, *6*(1), 368–403. https://doi.org/10.63125/c3g3sx44
[45]. Hammad, S., & Muhammad Mohiul, I. (2023). Geotechnical And Hydraulic Simulation Models for Slope Stability And Drainage Optimization In Rail Infrastructure Projects. *Review of Applied Science and Technology*, *2*(02), 01–37. https://doi.org/10.63125/jmx3p851

[46]. Hassan, R., Butler, M., O'Cearbhaill, R. E., Oh, D. Y., Johnson, M., Zikaras, K., Smalley, M., Ross, M., Tanyi, J. L., & Ghafoor, A. (2023). Mesothelin-targeting T cell receptor fusion construct cell therapy in refractory solid tumors: phase 1/2 trial interim results. *Nature medicine*, *29*(8), 2099-2109.

[47]. Heimonen, A., Nousiainen, K., Lassila, H., & Kaukiainen, A. (2023). Work-related head injury and industry sectors in Finland: causes and circumstances. *International archives of occupational and environmental health*, *96*(4), 577-586.

[48]. Hidayati, I., Tan, W., & Yamu, C. (2020). How gender differences and perceptions of safety shape urban mobility in Southeast Asia. *Transportation research part F: traffic psychology and behaviour*, *73*, 155-173.

[49]. Hu, Q., Cai, M., Mohabbati-Kalejahi, N., Mehdizadeh, A., Alamdar Yazdi, M. A., Vinel, A., Rigdon, S. E., Davis, K. C., & Megahed, F. M. (2020). A review of data analytic applications in road traffic safety. Part 2: Prescriptive modeling. *Sensors*, *20*(4), 1096.

[50]. Huang, C., Zhang, Z., Mao, B., & Yao, X. (2022). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, *4*(4), 799-819.

[51]. Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., & Zitnik, M. (2022). Artificial intelligence foundation for therapeutic science. *Nature chemical biology*, *18*(10), 1033-1036.

[52]. Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*, *12*(3), 1353.

[53]. Islam, S., Biswas, P. K., Saha, S., Sayem, A., & Khan, M. M. A. (2023). Occupational injuries and risk assessment among stone crushing industry workers: a cross-sectional study. *International archives of occupational and environmental health*, *96*(6), 903-917.

[54]. Jabed Hasan, T., & Waladur, R. (2023). AI-Driven Cybersecurity, IOT Networking, And Resilience Strategies For Industrial Control Systems: A Systematic Review For U.S. Critical Infrastructure Protection. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 144–176. https://doi.org/10.63125/mbyhj941

[55]. Jiang, H., Wang, M., Zhao, P., Xiao, Z., & Dustdar, S. (2021). A utility-aware general framework with quantifiable privacy preservation for destination prediction in LBSs. *Ieee/Acm Transactions on Networking*, *29*(5), 2228-2241.

[56]. Kaur, H., Wurzelbacher, S. J., Bushnell, P. T., Bertke, S., Meyers, A. R., Grosch, J. W., Naber, S. J., & Lampl, M. (2023). Occupational Injuries Among Construction Workers by Age and Related Economic Loss: Findings From Ohio Workers' Compensation, USA: 2007–2017. *Safety and health at work*, *14*(4), 406-414.

[57]. Khowaja, S. A., Dev, K., Qureshi, N. M. F., Khuwaja, P., & Foschini, L. (2022). Toward industrial private AI: A two-tier framework for data and model security. *IEEE Wireless Communications*, *29*(2), 76-83.

[58]. Knezek, G., Christensen, R., Smits, A., Tondeur, J., & Voogt, J. (2023). Strategies for developing digital competencies in teachers: Towards a multidimensional Synthesis of Qualitative Data (SQD) survey instrument. *Computers & Education*, *193*, 104674.

[59]. Kyung, M., Lee, S.-J., Dancu, C., & Hong, O. (2023). Underreporting of workers' injuries or illnesses and contributing factors: a systematic review. *BMC Public Health*, *23*(1), 558.

[60]. Lee, D., Lim, D., Park, J., Woo, S., Moon, Y., & Jung, A. (2024). Management Architecture With Multi-modal Ensemble AI Models for Worker Safety. *Safety and health at work*, *15*(3), 373-378.

[61]. Lee, Y.-C., Shariatfar, M., Rashidi, A., & Lee, H. W. (2020). Evidence-driven sound detection for prenotification and identification of construction safety hazards and accidents. *Automation in Construction*, *113*, 103127.

[62]. Leso, B. H., Cortimiglia, M. N., & Ghezzi, A. (2023). The contribution of organizational culture, structure, and leadership factors in the digital transformation of SMEs: a mixed-methods approach. *Cognition, Technology & Work*, *25*(1), 151-179.

[63]. Loosemore, M., Bridgeman, J., & Keast, R. (2020). Reintegrating ex-offenders into work through construction: A case study of cross-sector collaboration in social procurement. *Building research & information*, *48*(7), 731-746.

[64]. Ma, Y., Du, Y., Chen, Y., Gu, C., Jiang, T., Wei, G., & Zhou, J. (2020). Intrinsic Raman signal of polymer matrix induced quantitative multiphase SERS analysis based on stretched PDMS film with anchored Ag nanoparticles/Au nanowires. *Chemical Engineering Journal*, *381*, 122710.

[65]. Masud, R., & Md Sarwar Hossain, S. (2024). The Impact of Smart Materials And Fire-Resistant Structures On Safety In U.S. Public Infrastructure. *Journal of Sustainable Development and Policy*, *3*(03), 44-86. https://doi.org/10.63125/ygr1yk30

[66]. May, C. R., Albers, B., Bracher, M., Finch, T. L., Gilbert, A., Girling, M., Greenwood, K., MacFarlane, A., Mair, F. S., & May, C. M. (2022). Translational framework for implementation evaluation and research: a normalisation process theory coding manual for qualitative research and instrument development. *Implementation Science*, *17*(1), 19.

[67]. Md, K., & Sai Praveen, K. (2024). Hybrid Discrete-Event And Agent-Based Simulation Framework (H-DEABSF) For Dynamic Process Control In Smart Factories. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *4*(1), 72–96. https://doi.org/10.63125/wcqq7x08

[68]. Md Nahid, H., & Tahmina Akter Bhuya, M. (2024). An Empirical Study of Big Data–Enabled Predictive Analytics And Their Impact On Financial Forecasting And Market Decision-Making. *Review of Applied Science and Technology*, *3*(01), 143–182. https://doi.org/10.63125/1mjfqf10

[69]. Md Newaz, S., & Md Jahidul, I. (2024). AI-Powered Business Analytics For Smart Manufacturing And Supply Chain Resilience. *Review of Applied Science and Technology*, *3*(01), 183–220. https://doi.org/10.63125/va5gpg60

[70]. Md. Akbar, H. (2024). Computational Psychometrics and Digital Biomarker Modeling For Precision Mental Health Diagnostics. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 487–525. https://doi.org/10.63125/vg522x27

[71]. Md. Mosheur, R. (2025). AI-Driven Predictive Analytics Models For Enhancing Group Insurance Portfolio Performance And Risk Forecasting. *International Journal of Scientific Interdisciplinary Research*, *6*(2), 39–87. https://doi.org/10.63125/qh5qgk22

[72]. Md. Rabiul, K., & Khairul Alam, T. (2024). Impact Of IOT and Blockchain Integration On Real-Time Supply Chain Transparency. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 449–486. https://doi.org/10.63125/2yc6e230

[73]. Means, A. R., Kemp, C. G., Gwayi-Chore, M.-C., Gimbel, S., Soi, C., Sherr, K., Wagenaar, B. H., Wasserheit, J. N., & Weiner, B. J. (2020). Evaluating and optimizing the consolidated framework for implementation research (CFIR) for use in low-and middle-income countries: a systematic review. *Implementation Science*, *15*(1), 17.

[74]. Mehdizadeh, A., Cai, M., Hu, Q., Alamdar Yazdi, M. A., Mohabbati-Kalejahi, N., Vinel, A., Rigdon, S. E., Davis, K. C., & Megahed, F. M. (2020). A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling. *Sensors*, *20*(4), 1107.

[75]. Micheli, G. J., Farné, S., & Vitrano, G. (2022). A holistic view and evaluation of health and safety at work: enabling the assessment of the overall burden. *Safety science*, *156*, 105900.

[76]. Moreira, F. G., de Oliveira, C. P., & Farias, C. A. (2024). Workplace accidents and the probabilities of injuries occurring in the civil construction industry in Brazilian Amazon: A descriptive and inferential analysis. *Safety science*, *173*, 106449.

[77]. Navarro, C. L. A., Damen, J. A., van Smeden, M., Takada, T., Nijman, S. W., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., & Riley, R. D. (2023). Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *Journal of Clinical Epidemiology*, *154*, 8-22.

[78]. Nazar, W., Szymanowicz, S., Nazar, K., Kaufmann, D., Wabich, E., Braun-Dullaeus, R., & Daniłowicz-Szymanowicz, L. (2024). Artificial intelligence models in prediction of response to cardiac resynchronization therapy: a systematic review. *Heart Failure Reviews*, *29*(1), 133-150.

[79]. Neto, A. V. S., Camargo, J. B., Almeida, J. R., & Cugnasca, P. S. (2022). Safety assurance of artificial intelligence-based systems: A systematic literature review on the state of the art and guidelines for future work. *IEEE Access*, *10*, 130733-130770.

[80]. O'Donovan, R., & McAuliffe, E. (2020). A systematic review exploring the content and outcomes of interventions to improve psychological safety, speaking up and voice behaviour. *BMC health services research*, *20*(1), 101.

[81]. Oyelade, O. N., Ezugwu, A. E., Almutairi, M. S., Saha, A. K., Abualigah, L., & Chiroma, H. (2022). A generative adversarial network for synthetization of regions of interest based on digital mammograms. *Scientific reports*, *12*(1), 6166.

[82]. Ozturk, O. (2021). Bibliometric review of resource dependence theory literature: an overview. *Management Review Quarterly*, *71*(3), 525-552.

[83]. Paguay, M., Febres, J. D., & Valarezo, E. (2023). Occupational accidents in Ecuador: an approach from the construction and manufacturing industries. *Sustainability*, *15*(16), 12661.

[84]. Papa, L., Russo, P., Amerini, I., & Zhou, L. (2024). A survey on efficient vision transformers: algorithms, techniques, and performance benchmarking. *IEEE transactions on pattern analysis and machine intelligence*, *46*(12), 7682-7700.

[85]. Park, J., & Kang, D. (2024). Artificial intelligence and smart technologies in safety management: a comprehensive analysis across multiple industries. *Applied Sciences*, *14*(24), 11934.

[86]. Park, S., Park, C. Y., Lee, C., Han, S. H., Yun, S., & Lee, D.-E. (2022). Exploring inattentional blindness in failure of safety risk perception: Focusing on safety knowledge in construction industry. *Safety science*, *145*, 105518.

[87]. Peres, R. S., Jia, X., Lee, J., Sun, K., Colombo, A. W., & Barata, J. (2020). Industrial artificial intelligence in industry 4.0-systematic review, challenges and outlook. *IEEE Access*, *8*, 220121-220139.

[88]. Perifanis, N.-A., & Kitsios, F. (2023). Investigating the influence of artificial intelligence on business value in the digital era of strategy: A literature review. *Information*, *14*(2), 85.

[89]. Pishgar, M., Issa, S. F., Sietsema, M., Pratap, P., & Darabi, H. (2021). REDECA: a novel framework to review artificial intelligence and its applications in occupational safety and health. *International journal of environmental research and public health*, *18*(13), 6705.

[90]. Pittz, T. G., & Adler, T. R. (2023). Open strategy as a catalyst for innovation: Evidence from cross-sector social partnerships. *Journal of business research*, *160*, 113696.

[91]. Razavykia, A., Brusa, E., Delprete, C., & Yavari, R. (2020). An overview of additive manufacturing technologies—a review to technical synthesis in numerical study of selective laser melting. *Materials*, *13*(17), 3895.

[92]. Rickinson, M., Cirkony, C., Walsh, L., Gleeson, J., Salisbury, M., & Boaz, A. (2021). Insights from a cross-sector review on how to conceptualise the quality of use of research evidence. *Humanities and Social Sciences Communications*, *8*(1), 1-12.

[93]. Rifat, C., & Rebeka, S. (2023). The Role Of ERP-Integrated Decision Support Systems In Enhancing Efficiency And Coordination In Healthcare Logistics: A Quantitative Study. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 265–285. https://doi.org/10.63125/c7srk144

[94]. Rockström, J., Gupta, J., Qin, D., Lade, S. J., Abrams, J. F., Andersen, L. S., Armstrong McKay, D. I., Bai, X., Bala, G., & Bunn, S. E. (2023). Safe and just Earth system boundaries. *Nature*, *619*(7968), 102-111.

[95]. Sabuj Kumar, S. (2023). Integrating Industrial Engineering and Petroleum Systems With Linear Programming Model For Fuel Efficiency And Downtime Reduction. *Journal of Sustainable Development and Policy*, *2*(04), 108-139. https://doi.org/10.63125/v7d6a941

[96]. Sabuj Kumar, S. (2024). Petroleum Storage Tank Design and Inspection Using Finite Element Analysis Model For Ensuring Safety Reliability And Sustainability. *Review of Applied Science and Technology*, *3*(04), 94–127. https://doi.org/10.63125/a18zw719

[97]. Sabuj Kumar, S. (2025). AI Driven Predictive Maintenance In Petroleum And Power Systems Using Random Forest Regression Model For Reliability Engineering Framework. *American Journal of Scholarly Research and Innovation*, *4*(01), 363-391. https://doi.org/10.63125/477x5t65

[98]. Sai Praveen, K. (2024). AI-Enhanced Data Science Approaches For Optimizing User Engagement In U.S. Digital Marketing Campaigns. *Journal of Sustainable Development and Policy*, *3*(03), 01-43. https://doi.org/10.63125/65ebsn47

[99]. Salhab, W., Ameyed, D., Jaafar, F., & Mcheick, H. (2024). A systematic literature review on ai safety: Identifying trends, challenges and future directions. *IEEE Access*.

[100]. Sarker, I. H., Kayes, A., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big data*, *7*(1), 41.

[101]. Schaufeli, W. B., Desart, S., & De Witte, H. (2020). Burnout Assessment Tool (BAT)—development, validity, and reliability. *International journal of environmental research and public health*, *17*(24), 9495.

[102]. Seblova, D., Berggren, R., & Lövdén, M. (2020). Education and age-related decline in cognitive performance: Systematic review and meta-analysis of longitudinal cohort studies. *Ageing research reviews*, *58*, 101005.

[103]. Serradilla, O., Zugasti, E., Rodriguez, J., & Zurutuza, U. (2022). Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects. *Applied Intelligence*, *52*(10), 10934-10964.

[104]. Sharpe, K., Afshar, T., St-Hilaire, F., & McLeod, C. (2022). Return-to-work after work-related injury in the construction sector: a scoping review. *Journal of occupational rehabilitation*, *32*(4), 664-684.

[105]. Shoflul Azam, T., & Md. Al Amin, K. (2024). Quantitative Study on Machine Learning-Based Industrial Engineering Approaches For Reducing System Downtime In U.S. Manufacturing Plants. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 526–558. https://doi.org/10.63125/kr9r1r90

[106]. Smith, E. A., Cooper, N. J., Sutton, A. J., Abrams, K. R., & Hubbard, S. J. (2021). A review of the quantitative effectiveness evidence synthesis methods used in public health intervention guidelines. *BMC Public Health*, *21*(1), 278.

[107]. Soldatos, J., Kefalakis, N., Despotopoulou, A.-M., Bodin, U., Musumeci, A., Scandura, A., Aliprandi, C., Arabsolgar, D., & Colledani, M. (2021). A digital platform for cross-sector collaborative value networks in the circular economy. *Procedia Manufacturing*, *54*, 64-69.

[108]. Ta, V.-D., Liu, C.-M., & Tadesse, D. A. (2020). Portfolio optimization-based stock prediction using long-short term memory network in quantitative trading. *Applied Sciences*, *10*(2), 437.

[109]. Tambon, F., Laberge, G., An, L., Nikanjam, A., Mindom, P. S. N., Pequignot, Y., Khomh, F., Antoniol, G., Merlo, E., & Laviolette, F. (2022). How to certify machine learning based safety-critical systems? A systematic literature review. *Automated Software Engineering*, *29*(2), 38.

[110]. Toronto, C. E. (2020). Overview of the integrative review. *A step-by-step guide to conducting an integrative review*, 1-9.

[111]. Tselentis, D. I., Papadimitriou, E., & van Gelder, P. (2023). The usefulness of artificial intelligence for safety assessment of different transport modes. *Accident Analysis & Prevention*, *186*, 107034.

[112]. Ucar, A., Karakose, M., & Kırımça, N. (2024). Artificial intelligence for predictive maintenance applications: key components, trustworthiness, and future trends. *Applied Sciences*, *14*(2), 898.

[113]. van Nunen, K., Reniers, G., & Ponnet, K. (2022). Measuring safety culture using an integrative approach: The development of a comprehensive conceptual framework and an applied safety culture assessment instrument. *International journal of environmental research and public health*, *19*(20), 13602.

[114]. Viceconti, M., Pappalardo, F., Rodriguez, B., Horner, M., Bischoff, J., & Tshinanu, F. M. (2021). In silico trials: Verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products. *Methods*, *185*, 120-127.

[115]. Viswan, V., Shaffi, N., Mahmud, M., Subramanian, K., & Hajamohideen, F. (2024). Explainable artificial intelligence in Alzheimer's disease classification: A systematic review. *Cognitive Computation*, *16*(1), 1-44.

[116]. Wang, Y., Machado, A., & Telea, A. (2023). Quantitative and qualitative comparison of decision-map techniques for explaining classification models. *Algorithms*, *16*(9), 438.

[117]. Wu, H., Li, X., Sun, F., & Zhao, Y. (2022). A status review of volumetric positioning accuracy prediction theory and static accuracy design method for multi-axis CNC machine tools. *The International Journal of Advanced Manufacturing Technology*, *122*(5), 2139-2159.

[118]. Xu, X., Li, J., Zhu, Z., Zhao, L., Wang, H., Song, C., Chen, Y., Zhao, Q., Yang, J., & Pei, Y. (2024). A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis. *Bioengineering*, *11*(3), 219.

[119]. Yang, Z., Sun, L., Sun, Y., Dong, Y., & Wang, A. (2023). A conceptual model of home-based cardiac rehabilitation exercise adherence in patients with chronic heart failure: a constructivist grounded theory study. *Patient preference and adherence*, 851-860.

[120]. Yedulla, N. R., Battista, E. B., Koolmees, D. S., Montgomery, Z. A., & Day, C. S. (2022). Workplace-related musculoskeletal injury trends in the United States from 1992 to 2018. *Injury*, *53*(6), 1920-1926.

[121]. Yigitcanlar, T., Desouza, K. C., Butler, L., & Roozkhosh, F. (2020). Contributions and risks of artificial intelligence (AI) in building smarter cities: Insights from a systematic review of the literature. *Energies*, *13*(6), 1473.

[122]. Yuan, L., Li, J., Li, R., Lu, X., & Wu, D. (2020). Mapping the evaluation results between quantitative metrics and meta-synthesis from experts' judgements: evidence from the Supply Chain Management and Logistics journals ranking. *Soft Computing*, *24*(9), 6227-6243.

[123]. Zaheda, K. (2025a). AI-Driven Predictive Maintenance For Motor Drives In Smart Manufacturing A Scada-To-Edge Deployment Study. *American Journal of Interdisciplinary Studies*, *6*(1), 394-444. https://doi.org/10.63125/gc5x1886

[124]. Zaheda, K. (2025b). Hybrid Digital Twin and Monte Carlo Simulation For Reliability Of Electrified Manufacturing Lines With High Power Electronics. *International Journal of Scientific Interdisciplinary Research*, *6*(2), 143–194. https://doi.org/10.63125/db699z21

[125]. Zhang, C., Hu, M., Di Maio, F., Sprecher, B., Yang, X., & Tukker, A. (2022). An overview of the waste hierarchy framework for analyzing the circularity in construction and demolition waste management in Europe. *Science of the Total Environment*, *803*, 149892.

[126]. Zhang, X., & Mahadevan, S. (2020). Bayesian neural networks for flight trajectory prediction and safety assessment. *Decision Support Systems*, *131*, 113246.

[127]. Zulqarnain, F. N. U., & Subrato, S. (2023). Intelligent Climate Risk Modeling For Robust Energy Resilience And National Security. *Journal of Sustainable Development and Policy*, *2*(04), 218-256. https://doi.org/10.63125/jmer2r39