DATA-DRIVEN PREDICTION

Data-Driven Contractor Quality Performance Evaluation and Non-Conformance Prediction in Saudi Infrastructure Projects

Zulfiulla Khan

Published November 2026

Abstract

Infrastructure clients require reliable evidence that contractors can deliver conforming work consistently, not merely that individual defects are corrected before handover. Conventional performance reporting often aggregates inspection results, audit findings and non-conformance reports (NCRs) after problems have already affected productivity. This conceptual review examines how structured quality records, exposure-adjusted indicators and interpretable predictive models can support contractor evaluation in Saudi infrastructure projects. Drawing on targeted literature concerning contractor selection, construction performance measurement, defect classification and non-conformance analytics, it proposes an integrated framework separating historical performance assessment from prospective risk prediction. The framework links inspection and test plans, material approvals, laboratory results, corrective and preventive actions, work-package characteristics and independent verification through a common data architecture. It specifies candidate indicators, a time-aware prediction design, model evaluation criteria and a governance process for translating warnings into engineering interventions. A hypothetical example illustrates the interpretation of a risk estimate and an illustrative allocation of monitoring resources. Neither constitutes empirical evidence. Particular attention is given to differences in inspection intensity, reporting incentives, contractor attribution, data leakage and the contractual consequences of algorithmic assessments. The proposed approach supports the objectives of Saudi Vision 2030 through more transparent infrastructure delivery, reduced avoidable resource consumption and improved accountability. Prospective validation on independently governed Saudi project datasets remains necessary before operational adoption. The contribution is a practical research framework for connecting contractor quality assurance with preventive decision-making without replacing contractual acceptance procedures or professional engineering judgment.

Keywords: contractor performance; non-conformance prediction; infrastructure quality; machine learning; quality governance; Saudi Arabia; Vision 2030

1 Introduction

Infrastructure quality is produced through thousands of decisions that are distributed across organisations, trades and locations. A contractor may comply with an approved method statement on one work front while facing material shortages, revised drawings or inadequate supervision on another. The resulting variation is rarely captured adequately by a single monthly quality percentage. For an infrastructure owner responsible for multiple construction packages, the challenge is to distinguish ordinary variation from deteriorating control before defective work becomes concealed, expensive to repair or difficult to attribute. A useful contractor assessment system therefore needs to describe both the condition of completed work and the reliability of the processes that produce it.

Saudi Arabia provides a particularly relevant context for this problem. Industrial cities, road networks, utility corridors and factory developments involve repeated activities across geographically separated sites. Contractors, consultants and independent laboratories generate records that could support comparative analysis, but those records may use incompatible terminology, inconsistent identifiers or different inspection frequencies. Local contractor assessment research has emphasized the need for selection methods that extend beyond price [1–3]. This paper extends that concern into the execution phase, where evidence about actual workmanship and the effectiveness of corrective action can inform continuing oversight rather than merely initial procurement decisions.

Quality management traditionally relies on approved project quality plans, inspection and test plans (ITPs), material submittals, audits and records of non-conforming work. These remain essential contractual controls. Their limitations arise when information is treated as administrative evidence rather than as a resource for identifying patterns. A weekly register may show that ten NCRs have been closed without revealing whether the same defect recurs across successive lots. A contractor with few recorded NCRs may have superior workmanship, limited inspection coverage or a culture that discourages reporting. Data-driven evaluation must address these alternative explanations before attaching meaning to numerical comparisons.

The research problem has two distinct components. The first is retrospective measurement: how reliably has a contractor met requirements over a defined period and comparable scope? The second is prospective inference: given information available today, which upcoming work packages warrant additional preventive attention? The former concerns documented performance, whereas the latter estimates an uncertain future event. Mixing these components into a single opaque ranking makes it difficult to interpret deterioration or challenge incorrect classifications. The framework proposed here retains separate outputs and introduces a documented decision process linking them to practical oversight.

Recent construction analytics research includes work on predicting the consequences of non-conformances, extracting root-cause information from reports and classifying quality-defect descriptions [4–6]. Such studies indicate useful technical directions but do not establish that a model trained elsewhere can be transferred directly to Saudi infrastructure. Differences in specifications, construction methods, reporting systems and project delivery arrangements require local calibration. This review consequently emphasizes the conditions necessary for responsible implementation rather than claiming a universal predictive accuracy or a demonstrated reduction in defects.

2 Aim, objectives and research questions

The aim is to develop a technically defensible conceptual framework for evaluating contractor quality performance and predicting future non-conformances in Saudi infrastructure projects, while preserving independent verification and contractual accountability.

The objectives are to (1) identify appropriate leading and lagging indicators for contractor quality assessment; (2) establish a structured data model linking inspections, tests, materials and corrective actions; (3) formulate a prospective prediction task with explicit temporal and exposure boundaries; (4) define validation, interpretation and intervention procedures; and (5) examine how responsible adoption could support durable infrastructure and efficient resource use under Saudi Vision 2030.

The corresponding research questions are: How can contractor performance be measured without rewarding under-reporting or penalising intensive inspection? Which routinely collected records can

provide actionable early warning signals? How should predictive models be evaluated across different contractors, projects and construction stages? What governance arrangements are needed to ensure that model outputs remain advisory, explainable and contestable? How can a preventive quality system contribute to sustainability without making unsupported environmental claims?

3 Review methodology

This manuscript is a targeted narrative synthesis and conceptual framework, not a completed systematic review or a meta-analysis. The literature basis consists of publicly discoverable publication records and accessible summaries relating to Saudi contractor evaluation, project-level performance assessment, construction defect analytics and non-conformance management. The principal publication window is 2020–2025. The conceptual design also draws on standard quality assurance and statistical modelling principles. No employer records, interviews, field measurements or proprietary contractor datasets were supplied for the study.

The search terms used to locate relevant records combined construction quality, contractor performance, Saudi Arabia, non-conformance, rework, machine learning and defect prediction. For a subsequent reproducible systematic review, searches should be repeated in Scopus, Web of Science and engineering-specific indexes using documented search dates and database-specific syntax. A suitable query would combine contractor OR subcontractor with quality OR nonconformance OR defect and prediction OR performance OR inspection, followed by a Saudi or infrastructure contextual filter where appropriate. The broader international evidence should be retained when it addresses a transferable method rather than a location-specific outcome.

Proposed eligibility criteria include peer-reviewed engineering or construction management studies that define performance indicators, analyse quality records, examine contractor selection, or explain a predictive method relevant to construction defects. Studies concerned solely with occupational safety or financial distress should be included only when they clarify shared data or governance challenges. Duplicate reports should be consolidated, and purely promotional material excluded from the evidence synthesis. Full-text eligibility, duplicate screening and study-quality ratings have not been completed; therefore no screening counts or claims of comprehensive coverage are presented.

Evidence should ultimately be extracted into a matrix recording project type, country, sample unit, outcome definition, predictor timing, validation design, reported uncertainty and limitations. A study predicting rework cost after NCR issuance answers a different question from one predicting whether a future inspection will fail. Similarly, a text classification task cannot demonstrate that future defects were prevented. These distinctions govern the thematic synthesis below. The framework is an analytical proposal requiring prospective testing, and the cited studies are used to motivate components rather than to imply that the integrated system has already been validated.

4 Contractor quality performance as a measurement problem

4.1 From compliance records to comparable indicators

The first design requirement is to define the denominator for every indicator. A raw NCR count is generally unsuitable for comparison because contractors execute different volumes of work and are subject to different surveillance programmes. A rate per eligible inspection lot may be useful when inspection definitions are stable; a rate per installed quantity may be preferable for repetitive

earthworks, paving or pipe installation. Neither is automatically fair across trades. The measurement unit should therefore be attached to a clearly identified activity and specification family. Comparing like with like is more defensible than attempting to correct all heterogeneity through a single weighted score.

First-pass acceptance measures the proportion of eligible inspections accepted without corrective work or resubmission. It distinguishes right-first-time production from eventual acceptance following repeated intervention. Nevertheless, a first-pass result may be influenced by inspector availability, incomplete submission documentation or inspection cancellation. A data dictionary should differentiate technical rejection, administrative return and cancelled inspection. Otherwise, improvements in document handling could be mistaken for improvements in workmanship. The records must also identify repeat submissions so that a corrected lot does not enter the denominator as an independent new installation.

Corrective action performance is equally important. Median closure duration is easy to communicate but can conceal long-standing severe cases. Reporting the distribution of open-case age, stratified by severity, offers additional insight. An overdue case should be evaluated against its approved due date, with extensions recorded rather than overwritten. Closure requires verification of the physical correction and, where necessary, evidence that the underlying cause has been addressed. A rapid administrative closure without such evidence should not increase the contractor's performance score.

A recurring defect offers a stronger signal of ineffective preventive action than a single isolated occurrence. Recurrence should be defined using a combination of failure mode, activity, responsible work package and a predetermined look-back window. A generic label such as poor workmanship is too broad to support reliable linkage. Conversely, highly specific free-text descriptions may obscure common mechanisms. Standardized classifications coupled with original narratives allow analysts to trace related events without suppressing engineering detail. Table 1 provides a candidate measurement dictionary, not a prescribed contractual rating system.

Table 1. Proposed contractor quality indicators and interpretive safeguards.

Indicator	Operational definition	Interpretation / safeguard
First-pass acceptance	Accepted first submissions / eligible first submissions	Exclude administrative returns; stratify by activity
NCR incidence	Distinct verified NCR events / eligible lots	Report inspection coverage and severity
Recurrence	Repeated verified failure modes / eligible follow-up lots	Use predefined look-back and activity codes
Overdue actions	Open actions past approved due date / open actions	Retain extension history
Traceability	Lots with complete required material/test linkage / eligible lots	Check completeness and authenticity
Inspection coverage	Completed required inspections / due inspections	Never reward reduced surveillance

4.2 Exposure, case mix and surveillance bias

A contractor executing complex utility crossings or deep excavations may encounter greater inherent technical risk than one working on repetitive surface improvements. Fair evaluation requires a

baseline that accounts for these differences. One option is to compare performance within activity strata and aggregate the results using a common portfolio distribution. Another is to estimate expected failures from a model that includes relevant complexity factors and compare observed outcomes with expectations. Both approaches require sufficient data. Where evidence is sparse, the system should report uncertainty rather than imply precision through decimal-heavy rankings.

Inspection intensity creates another measurement challenge. Enhanced surveillance can initially increase recorded defects because more deviations are detected. Penalizing this increase could encourage contractors to restrict access or avoid internal reporting. A balanced assessment should therefore display completed inspection coverage, internal self-reports and externally detected cases alongside failure rates. Verified self-identification before concealment may demonstrate effective internal control even when the number of reported problems rises. The aim is to reinforce transparent detection and timely correction, not to create a superficially clean register.

An important distinction exists between a defective outcome and responsibility for the outcome. A failed test may follow an incorrect client instruction, a supplier non-conformance, a design ambiguity or a contractor execution error. Preliminary records should not automatically assign contractual liability. Attribution fields should distinguish the reporting organisation, executing party, suspected cause and formally determined responsibility. Assessments used for procurement or contractual enforcement should rely on the verified attribution rather than an unreviewed algorithmic label. Disputed cases should be visible separately with a documented route for resolution.

4.3 Combining scores without concealing critical failures

A composite score can support management communication if its components, weights and limitations are transparent. For example, normalised measures of first-pass acceptance, material traceability, corrective action effectiveness and inspection coverage could be combined as a weighted mean. Weights should be agreed through a documented stakeholder process and tested for sensitivity. A change in ranking caused by a modest adjustment of weights signals that differences are not robust. The score should always be accompanied by its component measures and sample sizes so that users can understand the reason for apparent improvement or decline.

Certain failures cannot be offset by strong performance elsewhere. An unresolved structural deficiency or a missing mandatory test may constitute a contractual hold regardless of the aggregate score. Such non-compensatory conditions should operate as separate rules rather than as arbitrary negative points in a weighted index. A contractor may have excellent documentation and a high acceptance rate but still require immediate intervention on a specific safety-critical lot. The proposed system therefore separates the portfolio score from mandatory escalation conditions and prospective risk estimates.

5 Data architecture and preparation

5.1 Establishing a common quality record

A linked data architecture is essential for meaningful prediction. Each physical work lot should have a persistent identifier connecting location, work breakdown structure, contractor, activity, specification revision, inspection request, material batch and laboratory results. The relationship between the lot and individual NCRs may be one-to-many, and a single NCR may affect several lots. A relational design

that preserves these relationships prevents duplication during aggregation. A spatial reference can also help distinguish repeated issues along a utility alignment or across adjoining pavement sections.

The data architecture should retain both event time and record-entry time. An inspection may occur on Monday but be entered into the system on Wednesday. If a model is intended to forecast risks on Tuesday, the Wednesday record cannot be treated as available information merely because the inspection date precedes the forecast. This distinction is particularly important when corrective actions or final root-cause determinations are recorded retrospectively. Historical feature extraction should reconstruct the actual information state at each prediction cutoff. Without it, model evaluation may produce unrealistically optimistic results.

Laboratory results require explicit provenance. A numerical strength or density result should be linked to the test method, applicable specification, sample identifier, testing date and calibration status where relevant. A value that falls below a specified limit is not interchangeable with an unverified or missing test. Missing results may reflect laboratory turnaround time, incomplete documentation or failure to take the required sample. Each possibility has different operational implications. The system should preserve these distinctions and make missingness visible rather than silently replacing unavailable values with averages.

5.2 Harmonisation and quality control

Data integration should start with a small, stable vocabulary agreed across the participating organisations. Activity codes, defect categories and severity levels should be mapped to a controlled dictionary while retaining the original entry. Mapping decisions need version control because a change in classification practice can create artificial trends. The records should also distinguish observations, formal NCRs, rejected inspections and audit findings. Combining these event types without qualification may overstate the number of failures when several records describe the same physical condition.

Automated validation rules can detect impossible dates, duplicate identifiers, missing locations and inconsistent units. They should be complemented by periodic manual review of a stratified sample. A data steward may resolve formatting issues, but technical judgments about defect severity or contractual disposition belong to authorized quality personnel. Corrections must preserve a change history. The goal is not to clean away inconvenient evidence but to ensure that the analytical record accurately reflects the underlying project events.

Textual descriptions contain potentially valuable details about workmanship and process deviations. Research on interpretable defect report classification and root-cause analysis suggests that natural language methods can help organize these records [5, 6]. A practical system might use supervised classification to propose a defect category, followed by human confirmation for ambiguous or severe cases. It should accommodate English and Arabic descriptions, abbreviations and trade-specific vocabulary. Translation or normalization must not remove negation, units or qualifiers that change engineering meaning. Confidential correspondence should be excluded unless its use is explicitly authorized.

6 Predictive modelling design

6.1 Defining a prospective outcome

The proposed prediction unit is an eligible upcoming inspection lot. At a specified cutoff, the system estimates the probability that the lot will experience a verified technical non-conformance within a defined horizon, such as the next inspection cycle or seven days. The horizon must reflect the time available for preventive action. Outcomes should be linked to independent inspection or verified test results and should distinguish technical failures from incomplete paperwork. Cases with insufficient follow-up should not be treated as successful outcomes merely because no report has yet appeared.

A separate model could estimate the likelihood of recurrence after corrective action or the expected cost of rework once an NCR has been issued. Published work on interpretable prediction of non-conformance cost impacts provides a relevant precedent for the latter task [4]. However, such a model operates after discovery of a deviation and should not be described as forecasting its initial occurrence. Maintaining separate target definitions prevents inappropriate use of post-event information and helps managers select the intervention that matches the decision stage.

Candidate predictors include prior acceptance rates within the same activity, recent recurrence, overdue corrective actions, material approval status, previous laboratory results available before the cutoff, changes in work-front supervision and relevant environmental conditions. Features derived from final NCR disposition, later test results or closure records must be excluded when they were not available at prediction time. The model should avoid using personally identifying information or irrelevant attributes as shortcuts for contractor quality. Any variables associated with contractor identity require particular scrutiny because they may encourage memorization instead of transferable risk learning.

6.2 Candidate algorithms and interpretation

Regularised logistic regression provides a useful baseline for binary prediction. It produces an estimated probability and permits relatively direct examination of the relationship between predictors and the outcome, subject to the limitations of the model specification. Tree-based ensembles can capture nonlinear interactions, such as the combined effect of a recently changed material source and unresolved corrective actions. Their greater flexibility may be beneficial but also increases the need for careful validation and interpretation. No algorithm should be chosen solely because it reports high training accuracy.

Hierarchical models may be appropriate when observations are clustered by contractor, site or work package. Partial pooling can stabilize estimates for contractors with limited records while retaining evidence of meaningful differences. This is especially important when a new contractor enters a programme and has no project-specific history. Predictions for such contractors should explicitly reflect the limited evidence rather than inherit a misleadingly confident average. A rule-based preventive checklist remains a reasonable fallback when historical data are insufficient for a robust model.

Interpretability tools can help explain why a particular lot is assigned elevated risk. For an ensemble model, feature attribution may identify contributions from a recent sequence of rejections, overdue actions or incomplete material traceability. These explanations describe the model's behavior, not proven physical causes. Correlated variables may share or exchange attribution, and a strong statistical association may arise from reporting practices. A quality engineer should investigate the

underlying evidence before selecting an intervention. Root-cause analysis still requires engineering reasoning and, where appropriate, direct examination of work and materials.

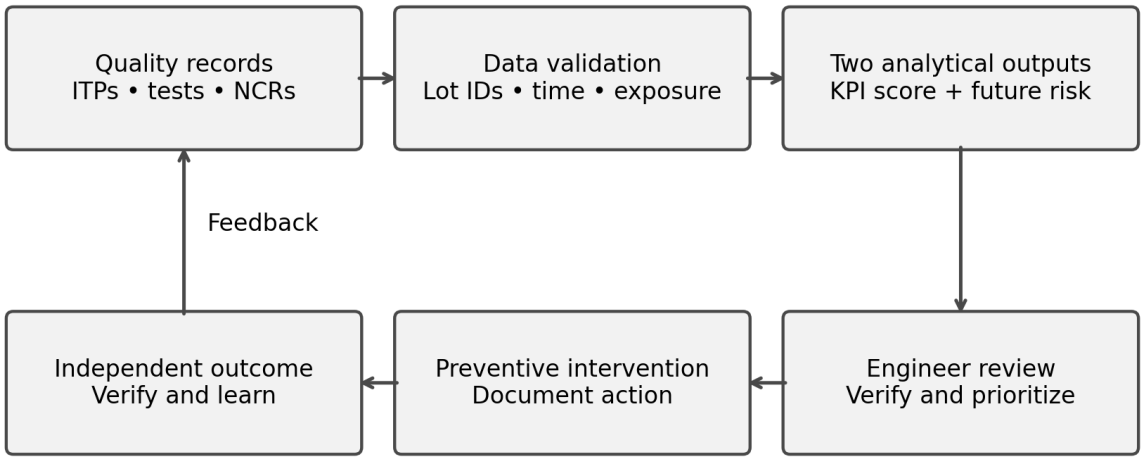

Figure 1. Proposed contractor assessment and non-conformance prediction workflow (conceptual).

6.3 Probability calibration and decision thresholds

A predicted probability is operationally useful only if it has an interpretable relationship with observed event frequency. Calibration should be assessed using reliability plots and appropriate probability scoring rules. If lots assigned approximately thirty percent risk experience verified failures substantially less often, the model may trigger excessive interventions. Conversely, systematic underestimation can lead to missed opportunities. Calibration should be evaluated on held-out projects or future periods, and recalibration should be considered when the specification, reporting practice or contractor population changes.

Thresholds should be selected in relation to the costs and consequences of the action. A low-cost document review may justify a lower threshold than an intrusive additional test. Critical work governed by mandatory hold points should remain subject to those requirements regardless of the predicted probability. A simple decision rule compares the expected avoidable loss from a potential failure with the incremental cost and effectiveness of a preventive intervention. Its inputs should be treated as uncertain and reviewed by engineering and commercial personnel, rather than automatically determining contractual action.

7 Validation and evaluation

A credible evaluation must preserve the structure of construction data. Randomly splitting individual inspection records may place neighboring lots, repeated submissions or the same contractor's nearly identical work in both training and testing sets. This creates dependence and can exaggerate generalization. A chronological split should train on earlier periods and test on later periods; a complementary grouped split should hold out entire projects or contractors. Where possible, a separate external project should be reserved for final assessment. Hyperparameter selection belongs within the training process, not the final test dataset.

Class imbalance is likely when verified technical failures represent a minority of eligible lots. Accuracy alone can be misleading because a model predicting universal acceptance may appear successful. Recall quantifies the proportion of actual failures identified, while precision describes how many alerts correspond to confirmed failures under a defined verification protocol. Precision-recall analysis should be interpreted in relation to the underlying event prevalence. Calibration, lead time and the number of investigations per prevented or detected failure are necessary complements to discrimination metrics. Table 2 summarizes the proposed evaluation design.

Table 2. Proposed model validation and operational evaluation.

Dimension	Measure	Design safeguard
Discrimination	Precision, recall, area under precision-recall curve	Report event prevalence and grouped test split
Calibration	Brier score and reliability plot	Use future or held-out project data
Actionability	Lead time before feasible correction ends	Record cutoff and intervention timing
Burden	Investigations per confirmed alert	Verify sampled non-alerted lots too
Fairness	Error rates by activity, contractor and site	Review exposure and case mix
Impact	Rework quantity, recurrence and inspection effort	Compare with standard practice in prospective pilot

The verification protocol must include some non-alerted lots. If only flagged work is subjected to additional checks, the true frequency of undetected problems cannot be estimated reliably. A stratified independent sampling design can provide evidence across predicted risk groups while maintaining normal contractual inspection coverage. The evaluation should also recognize that preventive intervention can change the eventual outcome: a high-risk lot that is corrected before inspection may pass because the alert was useful. Recording the intervention and its timing is essential for distinguishing predictive performance from preventive effectiveness.

A prospective pilot should begin in shadow mode, where predictions are generated without altering inspection decisions. This phase tests data availability, timing and calibration under operational conditions. After review, a controlled deployment can compare standard quality management with a supplementary risk-guided intervention process. Allocation should occur at an appropriate cluster level to minimize contamination between work fronts. Outcomes should include confirmed recurrence, rework quantities, inspection burden and any adverse effects. A pilot should not reduce mandatory verification requirements for the purpose of demonstrating efficiency.

7.1 Worked hypothetical example

Consider two comparable earthworks packages, each containing one hundred eligible inspection lots in a reporting period. Contractor A has eighty-eight first-pass acceptances, while contractor B has ninety-four. A recorded acceptance rate of eighty-eight versus ninety-four percent describes the observed sample, but does not establish a statistically or contractually meaningful difference. If contractor A received more complex work, a revised soil specification or more intensive verification, further investigation is required. The comparison should be accompanied by coverage, activity mix and confidence intervals rather than a categorical judgment of competence.

Suppose an upcoming lot is assigned a hypothetical probability of 0.30 for a verified technical failure. Assume, solely for illustration, that the avoidable loss associated with such a failure is SAR 20,000

and that a preventive review would avert half of that loss when a failure would otherwise occur. The expected avoidable loss is then SAR 3,000. If the review costs SAR 800, the simplified expected net benefit is SAR 2,200. These assumptions are invented to demonstrate decision logic; they are not measurements, model outputs or estimates from a Saudi contractor. Real decisions must account for uncertainty, potential delay, mandatory inspections and consequences that cannot reasonably be monetised.

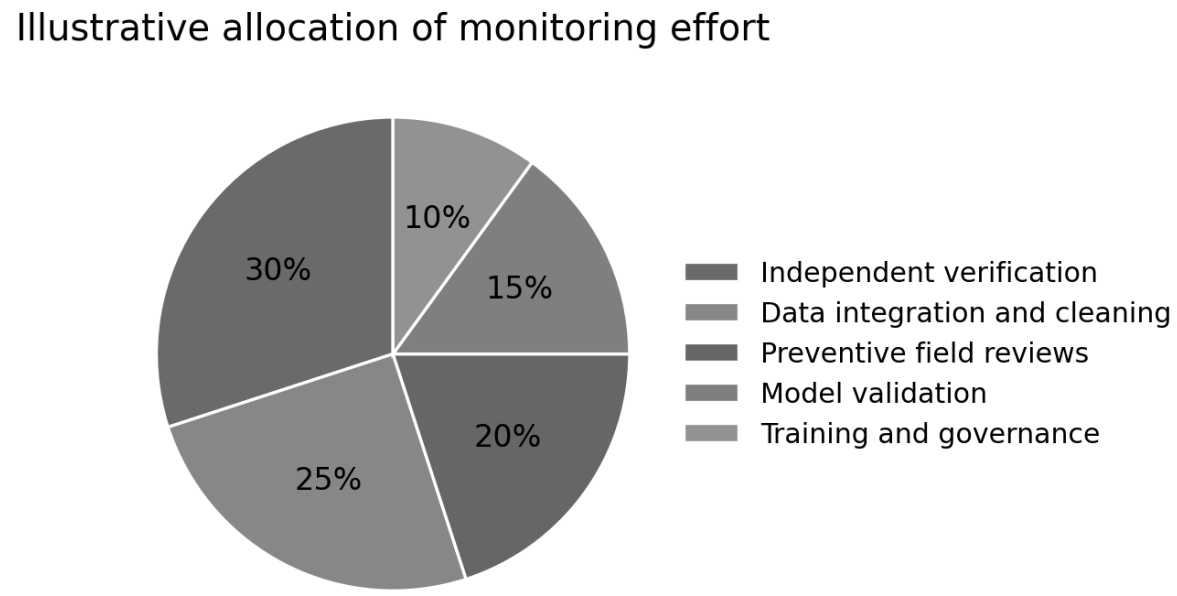

Figure 2. Hypothetical allocation of 100 monitoring-effort units; not observed contractor or project data.

8 Implementation in Saudi infrastructure projects

A phased implementation should prioritise work types with stable specifications and repeatable inspection units. Earthworks compaction, concrete placement, utility installation and asphalt paving offer plausible starting points because they produce structured testing records. However, the appropriate indicators differ. Compaction results depend on material and moisture conditions; concrete assessment may require strength results that become available after placement; utility installation includes dimensional, pressure and bedding checks; pavement quality involves multiple plant and field variables. The analytical system should respect these technical differences instead of imposing a single generic defect model.

In industrial city developments, multiple contractors may share access routes, utilities or approved suppliers. A data architecture that connects material batches and work locations can help identify whether similar deviations emerge across organizational boundaries. A recurring issue associated with a shared source should prompt a coordinated technical investigation rather than automatic attribution to every contractor. Consultant-led quality meetings can use aggregated trends to prioritise common preventive actions, while individual cases remain subject to established communication and contractual processes. This combination supports learning without obscuring responsibility.

The operational workflow should deliver alerts to the personnel capable of acting on them. An elevated risk associated with an upcoming concrete pour might require review of approved mix documentation, equipment readiness and inspection hold points. A recurring utility installation issue might trigger a targeted method-statement briefing or verification of bedding material before concealment. Each intervention should be linked to the alert and recorded with its rationale and result. The feedback allows the programme to determine whether the warning was actionable, redundant or based on a data error.

9 Governance, sustainability and Vision 2030

Saudi Vision 2030 provides a broad policy context for efficient, reliable infrastructure delivery. The framework's potential contribution is not the use of artificial intelligence in isolation, but the reduction of avoidable material loss, repeated construction activities and delays attributable to quality failures. These outcomes require measurement against an appropriate baseline. A reduction in recorded NCRs alone cannot establish sustainability benefits, because reporting may have changed. Physical quantities of replaced material, additional equipment hours and verified rework events offer a more direct basis for environmental and economic assessment.

Quality data also carry commercial and contractual consequences. Contractors should be informed of indicator definitions, data access rules and processes for correcting inaccuracies. A consultant or client should retain authority over formal acceptance and contractual decisions, with the model providing supplementary evidence only. Records must be access-controlled, and cross-project data sharing should be governed by permissions and confidentiality obligations. The submitting author should not use employer or client data without approval. This framework is designed around generic project workflows rather than unpublished company evidence.

A governance committee should periodically review model calibration, reporting incentives and the distribution of alerts across activity categories and contractors. If a model disproportionately flags a particular group, the explanation may involve genuine exposure differences, inadequate data or biased inspection practices. The committee should examine these alternatives and require remediation when the model produces unsupported conclusions. Automated contractor exclusion or commercial penalties based solely on predictions are inappropriate. Transparent review is particularly important where future procurement opportunities could be affected by performance ratings.

10 Research gaps and future research

The first research gap is the limited comparability of quality datasets across contractors and projects. A shared minimum data dictionary should be tested across multiple Saudi infrastructure programmes, with explicit definitions of inspection units, technical failures, corrective actions and attribution. Research should quantify how much apparent variation in contractor performance is explained by case mix and surveillance intensity. The results could inform a more equitable evaluation method, but data harmonisation must precede robust comparative inference.

A second gap concerns prospective evaluation. Retrospective classification of defect reports and prediction of already discovered non-conformance consequences are valuable but do not establish early warning effectiveness. Future research should investigate whether time-aware models provide meaningful lead time and whether preventive actions actually improve outcomes. A multi-project design with independently verified outcomes would help separate the model's contribution from

changes in management practices or contractor workload. The study protocol should specify how interrupted work, missing follow-up and interventions are handled.

A third gap involves the economic and environmental effects of preventive quality management. Research should link verified rework quantities to cost records and appropriate lifecycle inventory factors, with uncertainty ranges for material and energy inputs. It should avoid assuming that more inspections automatically reduce environmental impacts: additional sampling, travel and equipment use also consume resources. Comparative assessment should focus on net effects across the relevant construction process and, when feasible, subsequent asset performance.

11 Discussion and limitations

The proposed framework connects two traditionally separate activities: evaluating contractor performance and anticipating future quality failures. Its key distinction is that observed performance remains an auditable description of documented outcomes, whereas predictive risk is a probabilistic estimate used to prioritise attention. This separation prevents an uncertain warning from being treated as a confirmed failure. It also creates a coherent way to combine quality engineering, digital records and managerial decision support without changing the authority of contractual acceptance procedures.

Published Saudi contractor selection and performance studies [1–3] establish a relevant context for multidimensional assessment, while international construction analytics research [4–9] suggests opportunities for extracting actionable information from quality records. Nevertheless, the cited literature is heterogeneous in purpose and evaluation design. The present paper does not pool results, estimate a common effect size or assert that any particular machine learning method is superior across projects. The proposed integration represents a conceptual contribution rather than a validated empirical finding.

Limitations include the targeted rather than systematic literature search, incomplete full-text appraisal, absence of primary project data and lack of local model calibration. The illustrative example and resource allocation chart are deliberately hypothetical. No predictive model has been trained, no contractor has been ranked, and no measured cost savings or environmental benefits are reported. These restrictions should be retained in any submission until appropriate data and independent validation are available.

12 Conclusion

Data-driven contractor quality evaluation can become more informative when inspection outcomes, material records, corrective actions and activity exposure are linked within an auditable architecture. Prospective non-conformance prediction can complement these records by highlighting upcoming work that warrants additional preventive review. The proposed framework emphasizes temporal integrity, case-mix adjustment, calibration, independent verification and human accountability. Its application to Saudi infrastructure projects could support reliable delivery and reduced avoidable rework, consistent with Vision 2030 objectives. The next step is a governed multi-project validation programme that tests both predictive performance and the practical value of interventions. Until such evidence is available, the framework should be treated as a research proposal and decision-support design rather than a proven operational system.

References

[1] Development of a Performance Assessment Model for Contractors in Saudi Arabian Construction Projects. (2024). https://doi.org/10.1155/2024/8780539

[2] A Performance Quality Index to Assess Professional Conduct of Contractors at Sustainable Construction Projects in Saudi Arabia. (2023). Sustainability, 15(9), 7500. https://doi.org/10.3390/su15097500

[3] Development of a pre-qualification and selection framework for construction projects' contractors in Saudi Arabia. (2022). https://doi.org/10.1080/13467581.2022.2087657

[4] Predicting Cost Impacts of Nonconformances in Construction Projects Using Interpretable Machine Learning. (2024). Journal of Construction Engineering and Management, 150(1), 04023143. https://doi.org/10.1061/JCEMD4.COENG-13857

[5] Comprehensive Root Cause Analysis of Construction Defects Using Semisupervised Graph Representation Learning. (2023). Journal of Construction Engineering and Management. https://doi.org/10.1061/JCEMD4.COENG-13435

[6] Interpretable machine learning-based text classification method for construction quality defect reports. (2024). https://www.sciencedirect.com/science/article/pii/S2352710224008982

[7] AI-Enhanced Defect Identification in Construction Quality Prediction: Hybrid Model of Unsupervised and Supervised Machine Learning. (2023). Procedia Computer Science. https://doi.org/10.1016/j.procs.2023.12.066

[8] Enhancing Construction Performance: A Critical Review of Performance Measurement Practices at the Project Level. (2024). Buildings, 14(7), 1988. https://doi.org/10.3390/buildings14071988

[9] Non-conformances in infrastructure engineer-to-order production systems. (2025, online publication). https://doi.org/10.1080/00207543.2025.2505156