PETROLEUM ENGINEERING

Integration of Predictive Analytics and Quality Management Systems for Early Defect Detection in Oil and Gas Infrastructure Construction

Mohammad Faisal Khan

Published November 2026

Abstract

Predictive analytics can help oil and gas construction teams prioritize earlier inspections. This paper proposes a design that connects structured construction events, verified quality outcomes, risk scoring and inspector-led triage within an existing quality management system. It distinguishes forecasts of future nonconformity from automated screening of an image already acquired. A structured narrative review informs the selection of data, baseline methods, evaluation criteria and operational safeguards. An illustrative piping workflow and a numerical example show how lead time, inspection capacity and prevalence influence deployment. A prospective pilot is proposed to test operational value and detect unintended changes in inspection coverage. Its relevance to Saudi industrial development lies in traceable handover, development of inspection and analytics capability and avoidance of preventable late rework. The paper presents a conceptual framework rather than measured field performance.

Keywords: Predictive analytics; quality management systems; oil and gas construction; defect detection; weld inspection; Saudi Vision 2030

1 Introduction

Construction defects in oil and gas infrastructure can be straightforward to correct while an item remains accessible and expensive to address after testing, insulation or system handover. A piping spool can accumulate evidence from material receipt, fabrication, weld preparation, inspection and installation. These records often sit in different systems and become available at different times. Quality staff may discover a recurring pattern only after several items have progressed through the same work front. By then, the project can face additional access costs and repeated verification across interconnected systems.

Predictive analytics offers a way to prioritize earlier scrutiny. Its use is meaningful only when the project defines the decision that follows a signal. For instance, a score issued before weld examination might trigger a review of traceability and process records. An image screening tool used after an authorized radiograph is acquired might direct a qualified reviewer's attention to a region of interest. These applications have different data, outcomes and potential consequences. Combining them in one accuracy figure would obscure whether an earlier decision was possible.

The construction quality management system supplies the institutional setting. It defines the approved procedures, inspection responsibilities, nonconformity records and corrective action workflows. Analytics can improve the timing and coordination of these activities if it identifies when evidence is incomplete or when combinations of observations deserve human attention. Saudi industrial programs provide a setting in which improved technical capability and documented asset

quality are relevant goals. The paper considers how those goals can be investigated through project outcomes rather than assigning strategic value to the use of software alone.

1.1 Aim and objectives

The aim is to design an auditable integration of predictive analytics and construction quality management for early defect detection. The objectives are to define an actionable prediction target, specify the necessary data lineage and label controls, compare practical modelling and evaluation approaches, design a human review workflow and propose a field pilot. Piping fabrication provides the principal illustration, with extensions to civil and other mechanical work considered where the logic transfers.

1.2 Research questions

Which information available before a defined construction milestone can support an actionable quality signal? How should model outputs enter the existing quality management process and remain distinct from acceptance decisions? Which evaluation methods account for rare defects, changing inspection intensity and time constraints? What conditions determine whether performance at one work front can transfer to another?

2 Review methodology

This structured narrative review combines literature on industrial visual inspection, machine learning evaluation, data leakage and quality management with selected standards and industry references. Sources were chosen for their contribution to the proposed workflow. Studies of visual inspection help characterize image-based tasks, while work on calibration and validation addresses the reliability of numerical predictions. Technical standards are considered as sources of scope and acceptance governance, subject to the asset and contract. Results reported for public image benchmarks are used to understand methodological opportunities and restrictions.

The synthesis distinguishes four questions for each source group. What data are available at the decision time? Who establishes the outcome used in evaluation? What actions can the prediction change? What evidence is required for an authorized release? The resulting design was examined using a hypothetical spool fabrication sequence and a separate arithmetic exercise. The paper uses neither plant records nor identifiable worker data. It makes no claims about measured model accuracy, causal effects or the typical defect rate on Saudi projects.

3 Literature review and critical analysis

3.1 Forecasting and inspection screening

A process model forecasts the likelihood of a defined outcome from information available earlier in the workflow. It might estimate the probability of an initial examination rejection among welds that will undergo an eligible examination. Input information might include joint category, procedure version, recent equipment events and relevant precursor observations. The intended action is a timely review or additional verification allowed by the project. The output is conditional on how the population and outcome are defined.

An image screening model operates on an existing inspection artifact. It can highlight suspected features on an authorized image or prioritize qualified interpretation. Research on crack detection and industrial surface defects shows that learned representations can identify patterns under controlled conditions. Transfer is constrained by differences in image acquisition, illumination, geometry, material and annotation practices. A radiographic indication and a visible surface defect are different phenomena. Appropriate model evaluation must preserve the examination context and the acceptance process specific to each.

An anomaly detector can be useful when labelled defective examples are scarce. It learns what familiar observations look like and identifies departures. This can reveal changes in equipment, image quality or process conditions. Such departures can also be benign. An anomaly score has a different interpretation from a calibrated probability of rejection. The interface should tell inspectors which signal they are seeing and what evidence they should consult before acting.

3.2 Labels, measurement and visibility

A nonconformity report is an administrative record whose timing and classification can vary. Training on every report as though it represented a confirmed technical defect creates noisy labels. A report about an incomplete document and a rejection verified during examination need different outcome definitions. Repaired items also require event lineage so that an initial outcome is preserved separately from a later successful retest.

In selectively inspected populations, a missing rejection is weak evidence of a sound item. Items chosen for examination may already differ in criticality, process conditions or prior concern. A model trained on their outcomes estimates risk within the observed population unless the sampling process is addressed. A sensible initial pilot can restrict its target to a consistent examination population. Broader claims require a credible verification design, such as a controlled sample within the limits allowed by the owner's procedures.

The time of data availability is as important as the time of the event. A shift record may describe conditions that existed before examination but be entered after the result was known. Using that record as a feature during training would give the model information it would lack in operation. Data extraction should maintain both timestamps and reconstruct what was actually available when a score would have been produced.

4 Proposed integration framework

4.1 Define the prediction contract

Start with the joint, spool or component as the prediction unit and identify a decision time before the intended action. For the piping example, one model is scored when fabrication status reaches a specified stage and estimates the chance of rejection at the first eligible examination within a defined follow-up period. The target class, examination method, inclusion rules and censoring procedure are documented before training. If no eligible outcome is observed by the evaluation cutoff, the record remains unresolved for that target.

The prediction contract also identifies acceptable actions. An alert might lead to a traceability check, consultation with a supervisor or an additional authorized examination where justified. It does not convert a probability into a finding or authorize a change to an approved examination plan. If the

selected decision point leaves too little time to complete the review, the score may be technically accurate and operationally unusable. Recording this lead time makes the intended use testable.

Different asset types should be treated as different tasks when their inspection mechanisms or acceptance criteria differ materially. A single enterprise interface can show several models, but each model should retain its target, scope and threshold. This separation reduces the chance that a useful probability for one process is interpreted as directly comparable to a score for another.

4.2 Build a traceable data chain

The minimum record includes the item identifier, package, material or batch reference where relevant, procedure revision, applicable stage, inspection event, result and disposition. It should link to the original authorized record. For piping work, a joint identity should remain stable across scheduling, examination and repair. Where two systems use different identifiers, the mapping should have an owner and a controlled method for handling duplicates and revisions.

Feature definitions should specify their unit, time window and source. A rolling rate of recent first-pass rejections requires a denominator and a rule for when results enter the calculation. A blank equipment status field can mean that no event occurred, that the field is outside scope or that synchronization failed. Treating every blank as a zero introduces meaning that the source system may never have intended. A data quality monitor should distinguish those cases and report them by site or work front.

The data chain should also protect the integrity of the underlying quality records. Analytics receives an authorized copy or controlled view, while official approvals remain in the quality system. Raw evidence and subsequent interpretations have separate identifiers. Corrections retain lineage, including when a reviewer changed a classification. This allows later performance analysis to distinguish modelling error from a change in the reference outcome.

4.3 Establish modelling baselines

A useful baseline is the current inspection prioritization process, including rules already applied by supervisors. A second baseline is a simple statistical model using a limited set of stable features. A regularized logistic model can provide an interpretable starting point for structured records. More flexible models can represent interactions among features but need to demonstrate additional benefit under held-out conditions. A comparison based only on model complexity offers little evidence of practical usefulness.

When several image patches come from one radiograph or several views from the same joint, all related observations should stay in the same data partition. Otherwise the test set may contain near duplicates of training evidence. For process data, entire projects, crews or time blocks can help evaluate transfer, depending on the intended use. Calibration should be checked on a partition separate from the training observations, with the final held-out period preserved for evaluation.

4.4 Route scores into existing decisions

When a score exceeds the approved triage threshold and the data are in scope, create a review task with the item, current stage, model version, scoring time, decision deadline and a short explanation. The reviewer should be able to locate the source observations and mark whether the alert was useful, inaccurate, too late or caused by insufficient data. Alert handling should preserve the difference between acknowledging a task and closing a verified nonconformity.

The responsible inspector evaluates any technical finding under the applicable procedure. Engineering dispositions remain with the assigned authority. When a model or data service is unavailable, the existing inspection and test plan continues to govern. The operational record should show when the model was unavailable and which items fell back to ordinary review. This matters because an outage concentrated in a demanding period can bias the apparent performance of a pilot. Table 1 summarises how each analytics output maps to a quality management action.

Table 1. Analytics outputs and quality decisions.

Analytic output	When available	Quality management action
Process risk score	Before the first eligible examination	Review traceability and process records within the decision window.
Image screening flag	After acquisition of an authorized image	Direct a qualified reviewer to a region for interpretation.
Missing or unsuitable inputs	At scoring or screening	Route through ordinary inspection and record the reason.

Note: Authorized personnel retain responsibility for examination, acceptance and disposition.

5 Illustrative application and quantitative example

5.1 Piping spool fabrication

A fabrication yard prepares a batch of process piping spools for installation. The quality system records the material identity, joint preparation status and relevant examination plan. At the proposed decision time, a score flags a cluster of joints associated with recent incomplete process records and a change in work sequence. The inspector reviews the evidence and determines whether the issue is administrative or calls for attention under the approved plan. The system records the action and the time remaining before the intended examination or next release.

An item with incomplete identity mapping should be routed for data correction before risk scoring. If a later radiograph is screened for possible indications, that screen creates a separate task with its own model version and evidence. A qualified reviewer remains responsible for interpretation. Later, the first eligible result is linked to the original process score regardless of whether a flagged joint was repaired. This preserves the target for evaluation while documenting any intervention that may have changed what was ultimately observed.

5.2 Threshold and workload

Table 2 uses hypothetical examination outcomes for 1,000 eligible joints. The 100 alerts capture 28 of 40 rejections, giving 28 percent precision and 70 percent recall.

Table 2. Illustrative examination outcomes by alert status (hypothetical data).

Alert status	Initial rejection	No initial rejection	Total
Flagged	28	72	100
Unflagged	12	888	900
Total	40	960	1,000

If the team can review only 60 alerts before the decision window closes, the 100-alert threshold cannot be implemented as designed. A different threshold or scheduled review capacity might be needed. Performance for the top 60 items would require its own observed counts. One cannot infer that it captures a fixed proportion of the 28 rejections. The distribution of scores and the time at which items arrive determine which work can be reviewed.

An owner might use a simplified decision analysis with the expected cost of review, the consequence of a missed target and the proportion of cases for which early intervention is feasible. Any assumed costs should be separated from measured outcomes. The decision can also reserve capacity for critical items that require human attention regardless of score. A low predicted probability does not remove the inspection requirements or the consequences of being wrong.

6 Pilot evaluation

6.1 Technical evaluation

Technical evaluation should report precision and recall across thresholds, calibration at operationally relevant score ranges and the distribution of lead time. Precision describes the fraction of flagged records that meet the target outcome. Recall describes the fraction of observed positive outcomes captured by the alerts. For an imbalanced outcome, a high overall accuracy can coexist with few useful positive detections. Report absolute counts alongside rates, especially where a shift or package has limited volume.

Assess stability across asset class, procedure and work location where sample sizes allow. Small groups can produce volatile estimates. Confidence intervals or clustered resampling help express uncertainty when related observations share conditions. An out-of-time test addresses the immediate question of future performance under familiar conditions. An independent work front tests a stronger claim about transfer. A site-specific recalibration should be documented and evaluated before changes are propagated elsewhere.

Time-related features and labels deserve an independent audit. A small random sample should be reconstructed from original events to verify that the score used only contemporaneously available information. The audit should test duplicate identity, outcome coding and whether repaired items were retained as distinct events. The pilot should maintain a frozen evaluation snapshot so later corrections can be compared transparently.

6.2 Operational evaluation

Begin in shadow mode, where the model generates scores and recommendations but existing inspection decisions proceed as usual. Use this phase to verify data flow, latency and the plausibility of alerts. An active pilot can then introduce reviewer tasks in selected work fronts while comparable fronts continue with ordinary practice. Exposure, workload and technical scope should be monitored so any observed improvement can be interpreted.

Operational outcomes include time from actionable signal to review, fraction of alerts resolved within the decision window, confirmed early interventions and recurrence of the target failure mode. Escaped defects and examination coverage are balancing measures. A decline in observed rejections is ambiguous if the examination population changed. The evaluation should also record additional reviews per confirmed actionable case and inspect whether the system diverts attention from routine responsibilities.

Model monitoring continues after introduction. A change in procedure, equipment, input quality or examination practice can alter the score distribution or its relationship to outcomes. The team should assign responsibility for monitoring, change approval and rollback. Changes that affect what inspectors see must be versioned and communicated so old and new results can be interpreted in their proper context.

6.3 Separate label verification from intervention

When a reviewer responds to a process alert, the action can change the outcome the model was intended to predict. Suppose an inspector finds a preparation issue and the yard corrects it before the first eligible examination. The later examination may pass. Treating the event as an ordinary false alert would ignore the possibility that the intervention changed the outcome. Treating every corrected issue as a prevented rejection would overstate the benefit. The pilot should record the verified precursor, its severity and the exact action separately from the examination result.

A shadow period helps establish performance under existing practice. During active use, a predefined adjudication panel can classify which alerts led to substantiated action and whether the action preceded the relevant milestone. Reviewers can adjudicate a sample of items from both flagged and unflagged groups where allowed. Outcome definitions should remain stable, with a separate analysis of verified precursor corrections. This approach lets the project assess both prediction quality and the workflow's effect on construction decisions.

If the intended deployment changes which items are examined, define the evaluation population before altering any inspection allocation. A sample selected solely because of high scores cannot establish a reliable rate of missed defects among low-scoring items. Where the procedure permits a supplemental sample, the project can include random items alongside targeted reviews. The two samples should be distinguished in reporting so readers can interpret how detection opportunity was created.

6.4 Assess transfer under realistic conditions

The pilot should specify where a model is expected to operate without retraining. A model developed in one yard may be suitable for a later phase at that yard, provided the procedures, records and examination conditions are comparable. A different contractor may use another result coding system or record the same condition at a different point in the workflow. Apparent drift can therefore arise from changed measurement as well as changed construction performance.

A transfer assessment can begin with descriptive comparisons of feature distributions, missing fields and outcome definitions. Where a field cannot be mapped reliably, remove it from the deployment scope or test a model that does not rely on it. For images, document whether acquisition settings, equipment or annotation rules differ. The external evaluation should use the procedures and workload that reviewers will actually encounter, including the full range of ordinary cases.

The owner can also define an abstention policy. When important inputs are absent or a record lies outside the intended model population, the interface returns an explicit reason and directs the item through ordinary procedures. Monitor abstention by work front and time. A system that abstains primarily on complex work can appear highly accurate on the remaining cases while offering little assistance where scrutiny is most valuable. Coverage and error measures belong in the same report.

7 Discussion

Integration creates value when a model changes an action that is still possible and the outcome can be measured. Early process signals and image screening may complement each other, but their benefits arise at different points. A process signal can encourage a check before downstream work. An image screen can improve review efficiency after acquisition. Measuring each against its intended use helps managers decide where the investment is warranted.

Data stewardship can be a larger initial investment than algorithm development. A fabricator and owner may use different records to represent the same joint. Correcting the mapping can improve handover even if predictive performance is limited. That benefit should be documented separately. At the same time, collecting data about individual workers can raise fairness and organizational concerns. The pilot can begin with process, material and equipment features and review any later proposal for person-level use under the applicable governance.

For Saudi industrial projects, the most credible contribution is the ability to repeat and audit a quality intervention. Local inspectors, fabrication supervisors and data specialists can jointly define meaningful labels, review errors and decide where recalibration is necessary. Procurement can require access to model-independent evidence and exportable inspection histories. These capabilities make it easier for the owner to sustain the workflow after a contractor or technology provider changes.

A project can retain an export of scores, review decisions and verified outcomes independently of the vendor platform. This allows a later audit to reconstruct why a joint was reviewed and whether action occurred before examination. It also gives the owner a basis for evaluating a model update against the same cases and the original operating conditions.

7.1 Align technical standards and the project decision

Construction quality teams should confirm the governing code and edition from the project's design basis and contract. Process piping and pipeline welding may be governed by different documents, and requirements for a storage tank concern another asset class. The review workflow should retain the acceptance criterion actually used for the item. A data model that merges all technical rejections into a single class can hide meaningful differences in examination methods and thresholds.

Quality managers can map each proposed analytic action to an existing procedure step. The mapping should name the role allowed to request a check, document how results enter the official record and establish whether the action affects a release date. The inspection plan remains the source of prescribed examinations and acceptance criteria. Analytics should be assessed on how effectively it helps teams use the plan and respond to credible new information.

8 Limitations and future research

This paper presents a conceptual design informed by a narrative review. Public image datasets have different acquisition conditions and defect distributions from particular construction projects. The hypothetical results cannot establish predictive accuracy or economic benefit. Construction-phase interventions can also change the eventual label, complicating simple comparisons between active and shadow operation. A prospective design should document those interventions and preserve an interpretable comparison.

Future research should investigate which data features transfer between fabrication yards, how label quality changes under common reporting definitions and whether a controlled increase in early review reduces later rework. A multi-site study could compare structured risk models, inspection image support and a combined workflow using explicit decision times. Further work could assess whether lower-cost data quality improvements deliver greater near-term value in settings with few verified outcomes.

9 Conclusion

Predictive analytics can support early defect detection when its target, data availability and follow-up action are specified together. The proposed integration connects a time-bound prediction to an accountable quality review and preserves authorized inspection and acceptance procedures. It distinguishes forecasting from examination screening, addresses selective labels and sets evaluation criteria based on lead time and review capacity. A prospective pilot is needed to establish whether those mechanisms improve construction outcomes on Saudi oil and gas projects. Traceable evidence and consistent technical governance are the basis for testing that contribution.

References

[1] American Petroleum Institute (2020). Welded tanks for oil storage (API Standard 650, 13th ed.). *API*, Washington, DC. https://www.api.org/products-and-services/standards/important-standards-announcements/standard650

[2] American Petroleum Institute (2021). Welding of pipelines and related facilities (API Standard 1104, 22nd ed.). *API*, Washington, DC. https://www.apiwebstore.org/standards/1104

[3] American Society of Mechanical Engineers (2024). Process piping (ASME B31.3-2024). *ASME*, New York. https://www.asme.org/codes-standards/find-codes-standards/b31-3-process-piping

[4] Bergmann P, Fauser M, Sattlegger D, Steger C (2019). MVTec AD — A comprehensive real-world dataset for unsupervised anomaly detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9592–9600. https://doi.org/10.1109/CVPR.2019.00982

[5] Breiman L (2001). Random forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324

[6] Cha YJ, Choi W, Büyüköztürk O (2017). Deep learning-based crack damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, 32(5), 361–378. https://doi.org/10.1111/mice.12263

[7] Dung CV, Anh LD (2019). Autonomous concrete crack detection using deep fully convolutional neural network. *Automation in Construction*, 99, 52–58. https://doi.org/10.1016/j.autcon.2018.11.028

[8] Fawcett T (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[9] Gneiting T, Raftery AE (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. https://doi.org/10.1198/016214506000001437

[10] Guo C, Pleiss G, Sun Y, Weinberger KQ (2017). On calibration of modern neural networks. *Proceedings of Machine Learning Research*, 70, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

[11] He H, Garcia EA (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[12] International Organization for Standardization (2026). Quality management systems — Requirements (ISO 9001:2026). *ISO*, Geneva. https://www.iso.org/standard/9001

[13] Japkowicz N, Stephen S (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), 429–449. https://doi.org/10.3233/IDA-2002-6504

[14] Kapoor S, Narayanan A (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

[15] Kingdom of Saudi Arabia (2016). Saudi Vision 2030. https://www.vision2030.gov.sa/en/

[16] Kingdom of Saudi Arabia (n.d.). National Industrial Development and Logistics Program. https://www.vision2030.gov.sa/en/explore/programs/national-industrial-development-and-logistics-program

[17] Mery D, Riffo V, Zscherpel U, Mondragón G, Lillo I, Zuccar I, Lobel H, Carrasco M (2015). GDXray: The database of X-ray images for nondestructive testing. *Journal of Nondestructive Evaluation*, 34, 42. https://doi.org/10.1007/s10921-015-0315-7

[18] National Institute of Standards and Technology (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). *NIST*, Gaithersburg, MD. https://doi.org/10.6028/NIST.AI.100-1

[19] Ribeiro MT, Singh S, Guestrin C (2016). "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

[20] Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. https://doi.org/10.1111/ecog.02881

[21] Roth K, Pemula L, Zepeda J, Schölkopf B, Brox T, Gehler P (2022). Towards total recall in industrial anomaly detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14318–14328. https://doi.org/10.1109/CVPR52688.2022.01392

[22] Saito T, Rehmsmeier M (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[23] Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, Chaudhary V, Young M, Crespo JF, Dennison D (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28. https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems

[24] Tabernik D, Šela S, Skvarč J, Skočaj D (2020). Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31, 759–776. https://doi.org/10.1007/s10845-019-01476-x

[25] Wang G, Liao TW (2002). Automatic identification of different types of welding defects in radiographic images. *NDT & E International*, 35(8), 519–528. https://doi.org/10.1016/S0963-8695(02)00025-7