

A machine learning-enabled digital twin framework for risk-prioritized autonomous underwater vehicle inspection of subsea oil and gas pipeline networks

Devarakonda Harish Kumar¹, Endalkachew Mosisa², Hirpa G Lemu³
and Naol Dessalegn Dejene⁴

¹Department of Mechanical Engineering, Wollega University Institute of Technology, Nekemte, Oromia, Ethiopia

²Department of Mechanical Engineering, College of Engineering and Technology, Wollega University, Nekemte, Ethiopia

³Department of Mechanical and Structural Engineering and Materials Science, Faculty of Science and Technology, University of Stavanger, N-4036 Stavanger, Norway

⁴School of Mechanical Engineering, Pusan National University, Busan 46241, Republic of Korea

E-mail: dharishkumar03@gmail.com (corresponding author)

Abstract. Inspecting subsea oil and gas pipelines is expensive and autonomous underwater vehicles (AUVs) have limited mission endurance, yet most operators still inspect on fixed time-based schedules that spend the same effort on every segment irrespective of its condition or the severity of a possible failure. This paper presents an open, laptop-deployable framework that combines historical incident data, machine learning and a graph-based digital twin to direct an AUV to the pipeline segments whose failure would be most consequential. Using 7813 gas-transmission and hazardous-liquid incident records from the United States Pipeline and Hazardous Materials Safety Administration (2010–2025), three interpretable models—logistic regression, random forest and extreme gradient boosting—are trained with stratified cross-validation and synthetic minority over-sampling to predict high-consequence failures (those causing injury, fatality, fire or explosion). The best model reaches a cross-validated area under the receiver operating characteristic curve of 0.74, and Shapley additive explanations attribute risk mainly to the transported commodity, the pipeline system type and installation vintage. Segment risk scores are propagated onto a representative subsea network graph to form a digital twin, over which three AUV inspection policies are compared in Monte-Carlo simulation. The proposed risk-prioritized policy detects approximately twice as many high-consequence failures as fixed-schedule inspection at an equal budget, and matches current-practice detection using about 60% less inspection effort. The entire system runs on free software with no specialised hardware, making risk-based subsea inspection accessible to operators who cannot afford commercial digital-twin platforms.

Keywords: digital twin, machine learning, predictive maintenance, subsea pipelines, autonomous underwater vehicle

1 Introduction

Oil and gas transmission pipelines form one of the largest and most safety-critical infrastructure systems in the world. More than three and a half million kilometres of transmission pipeline are in operation globally, moving the majority of the crude oil and natural gas that is transported between production sites, processing facilities and markets. A substantial and growing fraction of this network is subsea: export trunklines, infield flowlines and risers that connect offshore wells, manifolds and platforms. These assets are difficult and costly to reach, and when they fail the consequences—environmental release into the marine environment, fire, explosion, and loss of production—are severe and highly visible.

Pipeline operators manage this risk through an activity known as integrity management, a central part of which is periodic inspection. Subsea inspection is performed by remotely operated vehicles (ROVs) and, increasingly, by autonomous underwater vehicles (AUVs) carrying multibeam sonar, high-definition imaging and other sensors. Such campaigns are expensive: published figures for subsea pipeline inspection lie in the range of roughly US\$8 000 to US\$25 000 per kilometre, so that a single survey of a moderate offshore field can cost several million dollars. Because inspection budgets are finite, every kilometre inspected on a low-risk segment is a kilometre of budget not spent where it matters most.

Two developments make it timely to reconsider how that budget is allocated. First, historical pipeline incident records are now openly available at large scale, which makes it possible to learn—rather than assume—which pipe attributes are associated with severe failures. Second, machine learning methods for structured data have matured to the point where such patterns can be extracted reliably and, crucially, explained. This paper brings these two developments together with the concept of a digital twin, a continuously updatable computational model of a physical asset, to answer a concrete operational question: given a limited AUV inspection budget, which subsea pipeline segments should be inspected first?

The contribution of this paper is a complete, integrated and openly reproducible framework that (i) trains and compares interpretable machine-learning models of high-consequence pipeline failure on real incident data; (ii) explains their predictions using a game-theoretic attribution method; (iii) embeds the resulting risk estimates into a graph-based digital twin of a subsea network; and (iv) uses that twin to drive and quantitatively benchmark an AUV inspection-allocation policy against the fixed-schedule practice used in industry today. The framework is deliberately built from free software and is light enough to run on an ordinary laptop, so that it is usable by smaller operators and by national oil companies in developing economies that cannot afford commercial digital-twin systems.

2 Problem definition

The framework targets five problems that operators recognise from daily practice.

Ageing assets under flat budgets. A large share of the installed pipeline base is between fifty and seventy years old. Older assets are more prone to corrosion and material degradation, but inspection budgets have not grown to match. A method that concentrates limited inspection effort on the segments that matter most directly addresses this gap.

Fixed schedules waste effort. Regulatory frameworks and long-standing practice inspect on fixed intervals. The cost per kilometre is similar whether a segment is high- or low-risk, so budget is routinely spent inspecting segments that were never likely to fail dangerously, while genuinely risky segments wait for their scheduled turn.

Static risk scoring. Risk-based inspection has been promoted for two decades, but the risk scores used in practice are typically computed from fixed engineering formulae that are revised only occasionally. They do not learn from new incident reports as those reports accumulate.

Commercial digital twins are out of reach. Digital-twin platforms from major vendors cost from hundreds of thousands to several million dollars per year and require expensive deployment. Smaller and emerging-market operators cannot justify this, leaving a real and underserved need for a lightweight, transparent alternative.

AUV deployment lacks risk-aware routing. Modern inspection AUVs are mechanically capable, but their missions are still planned by fixed human schedules rather than driven automatically by data-derived risk. Supplying an AUV with a ranked, risk-prioritized list of segments would make existing robotic assets markedly more effective.

3 Related work and research gap

Three streams of prior work are relevant. The first applies machine learning—random forests, support vector machines, neural networks and gradient boosting—to predict pipeline failures or corrosion rates [1, 2]. These studies are valuable but usually consider a single pipe or short segment, predict a single failure mode, and stop at the prediction step without connecting it to an inspection decision or a network-wide view. The second stream builds digital twins of individual high-value assets such as a specific riser or pump [3]; these produce useful models but rarely scale to an entire interconnected network. The third stream studies inspection scheduling and the routing of inspection tools and vehicles [4], but typically uses risk scores that are fixed in advance by engineers rather than learned from data.

No study found in the open literature does all of the following together: trains a machine-learning model on real, large-scale incident data; embeds the resulting risk into a network-wide, graph-based digital twin; and uses that twin to drive an inspection-allocation policy that is quantitatively benchmarked against industry-standard baselines, while remaining open-source and light enough for laptop deployment. Providing that integrated, reproducible framework, in the subsea AUV context, is the gap this paper fills.

4 Proposed framework

The framework has three integrated layers, shown in figure 1. The *data layer* ingests and prepares historical incident records and accepts new incidents over time. The *intelligence layer* trains and compares machine-learning models that estimate, for a segment with given attributes, the probability that a failure there would be high-consequence, and explains those estimates. The *digital-twin and decision layer* represents the pipeline network as an attributed graph, writes the model’s risk estimates onto its edges, and lets an AUV inspection agent traverse the graph under competing policies that are compared quantitatively.

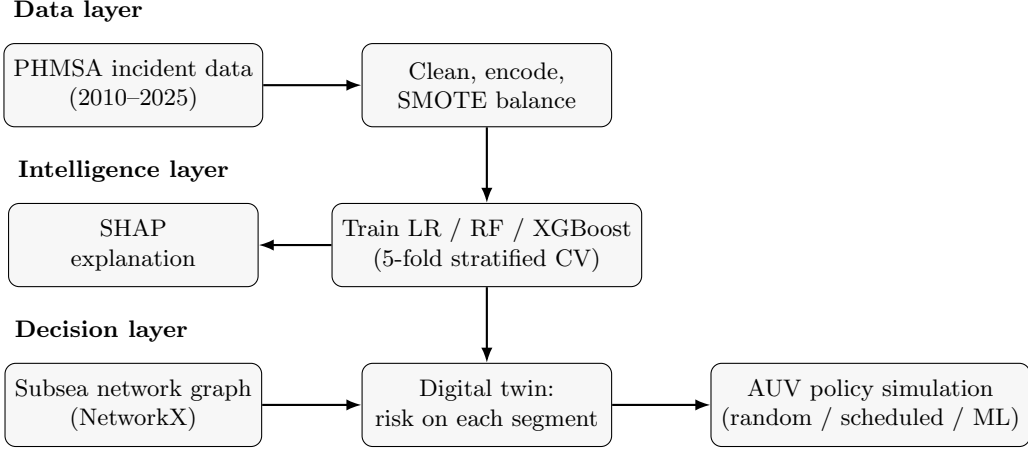


Figure 1: The three-layer framework: data preparation, model training and explanation, and the graph-based digital twin that drives AUV inspection allocation.

5 Materials and methods

5.1 Data source

The study uses the incident databases published by the United States Pipeline and Hazardous Materials Safety Administration (PHMSA) [5], the largest openly available and well-structured pipeline incident dataset. Two collections are combined: gas transmission and gathering incidents, and hazardous-liquid accidents, each covering January 2010 onward. After restricting to complete years (2010–2025), the combined dataset contains 7813 incident records (1984 gas-transmission and 5829 hazardous-liquid), of which 287 occurred offshore. Each record carries the installation year, pipe diameter, transported commodity, the part of the system involved, the failure cause, operating and consequence information. Although the framework targets subsea assets, the offshore-only records (287) are too few to train reliable models on their own; the full corpus is therefore used with an explicit offshore indicator as a feature, so that the model learns risk drivers—many of which, such as commodity, diameter and vintage, are common to onshore and offshore pipe—while still able to distinguish offshore behaviour. The learned drivers are then applied to score a subsea network.

Applying onshore-trained scores to a subsea network is justified by looking at which features actually drive the predictions. SHAP analysis (section 6.3) shows that commodity type, pipeline system classification and installation decade are the strongest predictors. These are properties of the pipe and its contents that affect failure consequence regardless of whether the pipe sits on the seabed or on land. The underlying physics is similar. The offshore indicator captures the systematic difference in consequence rates: offshore incidents show a lower high-consequence rate than onshore ones (figure 2), and that adjustment is estimated directly from the data. Operators with enough offshore-specific incident history could retrain on that subset for sharper scores; the full public corpus is used here as the best available proxy for a technology-demonstration study, with this limitation noted in section 5.7.

5.2 Feature engineering and target definition

The prediction target is a *high-consequence incident*, defined as an incident that caused a fatality, an injury, an ignition (fire) or an explosion. These are the outcomes integrity management most seeks to prevent, and they are rare: 5.7% of records in the dataset. Estimated repair cost, although recorded, is used only descriptively and never as a model feature, because it is known

only after a failure.

Framing the target this way makes the study a risk-prioritization exercise: the framework learns and ranks the *consequence component* of risk—the quantity most directly tied to the safety and environmental outcomes that subsea inspection exists to prevent—directing the AUV toward the segments whose failure would matter most. Modelling absolute failure likelihood, which requires pipeline-mileage exposure denominators, is identified as a complementary extension in section 7.

Using a severity-conditional model is appropriate here because the AUV scheduling problem is fundamentally about ranking, not about crossing a probability threshold. The vehicle needs to visit the most dangerous segments first, which means the relative ordering of risk scores matters more than their absolute values. Ranking by conditional severity puts inspection effort where failures would hurt most, which is the core safety objective. Adding failure frequency to the model—the full frequency-times-consequence risk product—would refine that ranking further, but requires exposure data that is not available in the public PHMSA records.

To ensure that the model can genuinely score an un-failed segment, only attributes knowable before a failure are used as features: pipe age (incident year minus installation year), pipe diameter, an offshore indicator, the system part involved, the transported commodity, the pipeline system type, and the installation decade. Failure-mechanism fields such as cause and release type, which describe the event after it has occurred, are deliberately excluded to avoid look-ahead leakage. Categorical variables are one-hot encoded and numerical variables are median-imputed and standardised.

5.3 Class balancing and model training

High-consequence incidents form a minority class (5.7%), so the synthetic minority over-sampling technique (SMOTE) [6] is applied to the training data. Critically, SMOTE is fitted *inside* each cross-validation fold, so that no synthetic sample can leak into the data used for validation, a common pitfall that would inflate reported performance.

Three interpretable models are trained and compared: logistic regression, a linear and highly interpretable baseline; random forest [7], an ensemble of decision trees; and XGBoost [8], a leading method for structured tabular data. Each model is evaluated with five-fold stratified cross-validation. For classification metrics (precision, recall, F_1) a threshold of 0.5 was used; for ranking and simulation tasks the raw predicted probability was used directly without any thresholding. Performance is measured by the area under the receiver operating characteristic curve (ROC-AUC), precision, recall and the Brier score. The hyperparameters for each model are listed in table 1; all other parameters were left at their library defaults.

5.4 Model explanation

Shapley additive explanations (SHAP) [9], an attribution method grounded in cooperative game theory, are applied to the selected model to quantify how each attribute pushes a prediction higher or lower. This matters in practice, because an integrity engineer will not act on a model that merely labels a segment risky without explaining why.

5.5 Digital twin and inspection simulation

A representative subsea production network is constructed as a graph using NetworkX [10]. The nodes are subsea junctions—a platform/riser base, six well-cluster manifolds, and the connected wellheads—and the edges are pipeline segments: trunklines from the riser base to the manifolds and branch flowlines from the manifolds to the wellheads, giving 43 nodes and 42 segments. Each

Table 1: Hyperparameters used for each model. All remaining parameters are at library defaults.

Model	Hyperparameter	Value
Logistic regression	Regularisation strength C	1.0
	Maximum iterations	2000
Random forest	Number of trees	400
	Maximum depth	Unlimited
	Minimum samples per leaf	2
XGBoost	Number of trees	400
	Maximum tree depth	5
	Learning rate	0.05
	Row sub-sample ratio	0.9
	Column sub-sample ratio	0.9

segment carries physical attributes—segment length, diameter, age, water depth and operating pressure, together with commodity and installation vintage—drawn from industry-plausible subsea ranges. The subset of attributes available in the public incident data (age, diameter, commodity, vintage, offshore class) is used by the trained model to score each segment’s high-consequence risk; the remaining attributes (length, water depth, operating pressure) are carried by the twin for routing and would enter the risk model directly where operator data provide them. Writing the risk scores onto the graph edges produces the digital twin.

The AUV is modelled as an agent that inspects segments under a limited budget. Three policies are compared: *random*, which inspects in random order and represents the worst case; *scheduled*, a fixed-interval systematic coverage that is risk-agnostic and represents current industry practice; and *ML risk-prioritized*, which inspects segments in descending order of predicted risk.

In each of 400 Monte-Carlo runs, the segments that suffer a high-consequence failure are sampled from a true per-segment propensity that the model estimates only imperfectly, so the ML policy is not clairvoyant. To model this estimation error, independent Gaussian noise ($\sigma = 1.0$ on the logit scale) was added to the model’s predicted scores to obtain the latent true propensity. The Spearman rank correlation between model scores and true propensity averaged 0.66 across runs. The noise level was set so the rank correlation is roughly proportionate to the model’s observed validation AUC of 0.74—a perfect ranker would have AUC 1.0 and correlation 1.0, while a random ranker would have AUC 0.5 and correlation 0. The sensitivity of results to this noise assumption is tested in section 6.6. Simulation settings: the horizon spans $T = 104$ inspection cycles; the budget is the fraction of the 42 segments inspected each cycle (headline results use 15%, roughly 6 segments per cycle); a failure is detected if its segment falls within the budget; mean time to detection (MTTD) is the average cycle number at which a failing segment is first reached, with undetected failures censored at $T = 104$. Policies are scored on detection rate as a function of budget, on mean time to detection, and on the budget the ML policy needs to match current-practice detection.

5.6 Computational environment and reproducibility

The entire framework runs in Python 3 using only free and open libraries: pandas and NumPy for data handling, scikit-learn [11] and imbalanced-learn for modelling and SMOTE, XGBoost for

gradient boosting, SHAP for explanation, NetworkX for the graph, and Matplotlib for figures. A single random seed is fixed throughout, and every figure and number in this paper is regenerated from the raw public data by the released code, so that the work runs unchanged on an ordinary laptop or on a free cloud notebook.

5.7 Scope and limitations

The boundaries of the study are stated plainly so that no reader is misled. The pipeline network used is representative rather than the property of a specific operator; the incident data are United States-based and may not perfectly reflect conditions elsewhere; the AUV is modelled as an agent traversing a graph rather than as a physical vehicle in a fluid-dynamics simulation; and the model estimates the conditional severity of a failure from historical incidents rather than the absolute frequency of failure, which would require exposure (mileage) denominators.

6 Results and discussion

6.1 Characterisation of the incident data

Figure 2 summarises the dataset. Equipment failure and corrosion are the most frequent causes, together accounting for over half of all incidents. The distribution of pipe age at failure is broad, with a median of 25 years and a long tail of much older assets. Offshore incidents, although a minority, show a distinctly lower high-consequence rate than onshore incidents, a plausible reflection of fewer people being present and of subsea releases dispersing in water rather than igniting.

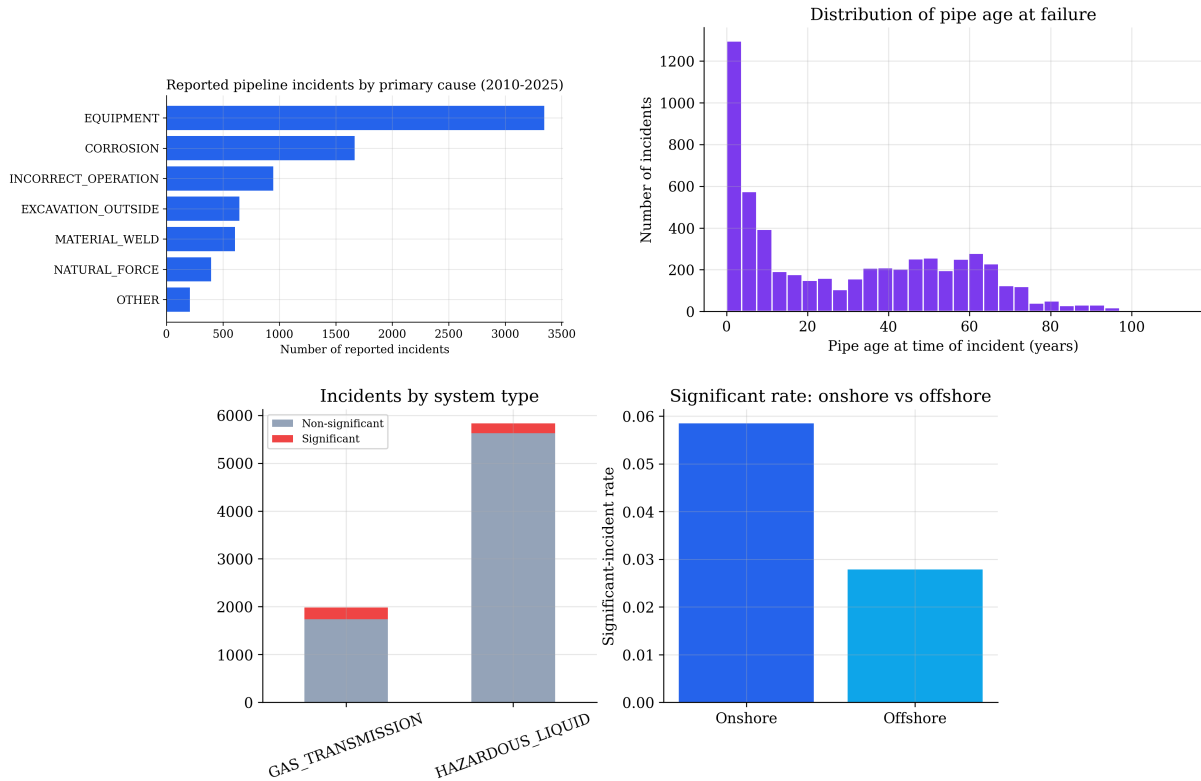


Figure 2: Descriptive statistics of the PHMSA incident data (2010–2025): incidents by primary cause (top left), pipe age at failure (top right), and the high-consequence rate by system type and by onshore/offshore location (bottom).

6.2 Model performance

Table 2 reports cross-validated performance for the three models, and figures 3 and 4 show the corresponding ROC, precision–recall and calibration curves. Logistic regression achieves the highest discrimination (ROC-AUC 0.735 ± 0.025) and recall, while the tree ensembles produce better-calibrated probabilities (lower Brier score). Because inspection prioritization is fundamentally a ranking problem—the AUV visits segments in risk order—the model with the strongest ranking ability is selected to drive the digital twin. An AUC of about 0.74 for predicting rare, high-consequence events from a small set of static attributes is a realistic and useful level of performance.

Offshore subset performance. When the trained model was evaluated only on the 287 offshore records (8 high-consequence positives; prevalence 2.8%), the ROC-AUC dropped to 0.634, still above the random baseline of 0.5, but below the full-corpus result of 0.735. At the default 0.5 threshold the model predicted no positives at all in the offshore subset, because the offshore prevalence is lower than the 5.7% full-corpus rate on which the threshold was calibrated. Lowering the threshold would recover offshore recall at the cost of precision. This performance gap is expected given so few offshore samples and confirms the limitation noted in section 5.7: retraining on offshore-specific data, where an operator can supply it, would sharpen the scores considerably. On the consequence scope, the outcome variable—injury, fatality, fire or explosion—covers the main safety consequences recorded by PHMSA. Environmental damage from subsea spills is not separately coded in the public data and would need a locally enriched dataset to incorporate. The Brier scores in table 2 need some context. A naive model that always predicts the marginal prevalence (5.7%) achieves a Brier score of $0.057 \times (1 - 0.057) \approx 0.054$. Logistic regression scored 0.205, well above that baseline, confirming its probability estimates are poorly calibrated. The ensemble models did better on calibration (RF 0.113, XGBoost 0.126) but still did not beat the naive baseline. None of this affects the operational conclusions: the models are used to *rank* segments, not to produce well-calibrated probabilities, and ROC-AUC is the right metric for that purpose.

Table 2: Five-fold cross-validated performance (mean $\pm$ standard deviation). The best value in each column is shown in bold.

Model	ROC-AUC	Precision	Recall	F_1	Brier
Logistic regression	0.735 $\pm$ 0.025	0.114	0.701	0.196	0.205
Random forest	0.707 $\pm$ 0.031	0.132	0.332	0.189	0.113
XGBoost	0.710 $\pm$ 0.019	0.122	0.420	0.188	0.126

6.3 Explanation of model decisions

Figure 5 shows the SHAP attributions. The strongest drivers of predicted high-consequence risk are the transported commodity—in particular highly volatile liquids—and the pipeline system type, followed by installation vintage and pipe age. This is physically coherent: failures of gas and highly volatile liquid lines are far more likely to ignite or explode than releases of ordinary liquids, which dominate the high-consequence outcomes. Notably, greater pipe age does not translate into higher predicted consequence once commodity and system type are accounted for: older lines fail more often through corrosion, but those failures are comparatively less likely to be high-consequence. This is an important, and slightly counter-intuitive, message for practice: prioritising inspection on age alone—as fixed schedules implicitly do—can misdirect

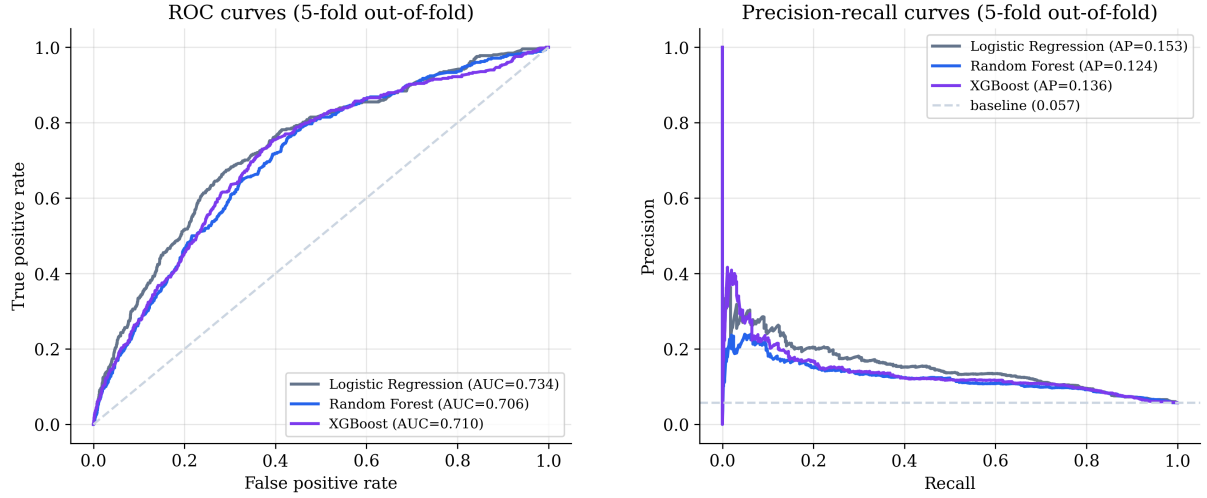


Figure 3: Receiver operating characteristic curves (left) and precision–recall curves (right) for the three models, computed from out-of-fold predictions.

effort, whereas a multi-factor risk model captures the commodity and system-type drivers that actually govern severity.

6.4 The digital twin

Figure 6 shows the digital twin: the representative subsea network with each segment coloured by its model-predicted high-consequence risk. The visualisation makes the risk distribution immediately legible—trunklines and flowlines carrying gas or highly volatile liquids stand out in red—and it is this risk field that the AUV inspection agent acts upon.

6.5 Comparison of inspection policies

Figure 7 and table 3 give the central operational result. At an inspection budget of 15% of segments per cycle, the ML risk-prioritized policy detects **24.3%** of high-consequence failures in time (95% CI: 8.3–41.7%), against **12.1%** for fixed-schedule inspection (95% CI: 0.0–27.3%) and **14.3%** for random inspection (95% CI: 0.0–30.8%). This is more than double the detection of current practice. Across the whole budget range the ML detection curve rises steeply and remains well above the two risk-agnostic policies, which track the diagonal expected of coverage that ignores risk. The ML policy also lowers the mean time to detection substantially (**35.9** cycles, 95% CI: 25.3–48.2, versus **51.4** cycles for fixed-schedule and **53.3** for random). The same advantage is visible in coverage efficiency (table 3): at a 15% budget the ML policy inspects 30% of the network’s total predicted risk, giving a coverage efficiency of 2.0, against about 0.76–0.95 for the risk-agnostic policies. Equivalently, the ML policy reaches the detection level of current-practice scheduling using only about 6% budget instead of 15%, a 60% reduction in inspection effort for the same safety outcome. **Note that this saving is expressed as a fraction of segments; if segment lengths vary widely or AUV transit time between distant segments is a significant cost, the saving should instead be expressed in pipeline kilometres and deployment cost.**

6.6 Sensitivity of policy advantages to rank-correlation assumption

The simulation uses a logit-scale noise of $\sigma = 1.0$, which produces a Spearman rank correlation of around 0.55–0.67 between model scores and true propensity across runs. To check how much

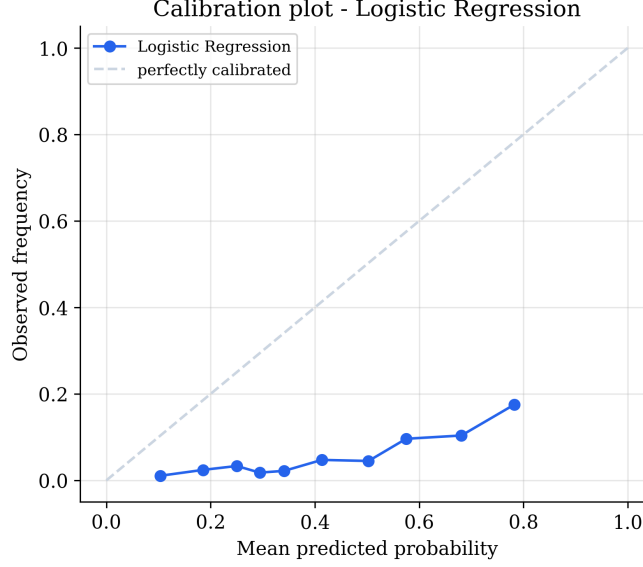


Figure 4: Calibration of the selected model: observed high-consequence frequency against mean predicted probability.

Table 3: Inspection-policy performance at a 15% per-cycle budget (400 Monte Carlo runs). Detection rate and mean time to detection (MTTD) are shown with 95% confidence intervals in brackets. Coverage efficiency is the fraction of the network’s total predicted risk inspected divided by the budget fraction (1.0 denotes risk-agnostic coverage).

Policy	Detection rate [95% CI]	MTTD cycles [95% CI]	Coverage efficiency
Random	14.3% [0.0, 30.8]	53.3 [39.4, 66.7]	0.95
Scheduled (fixed interval)	12.1% [0.0, 27.3]	51.4 [39.0, 62.9]	0.76
ML risk-prioritized	24.3% [8.3, 41.7]	35.9 [25.3, 48.2]	2.02

the results depend on this particular choice, the simulation was re-run at $\sigma \in \{0.6, 1.0, 1.4, 1.8\}$, giving rank correlations of roughly $\rho \in \{0.67, 0.55, 0.63, 0.20\}$. At 15% budget: for $\sigma = 0.6$ ($\rho \approx 0.67$), ML detected 24.5% versus 12.6% for scheduled; for $\sigma = 1.0$ ($\rho \approx 0.55$), ML detected 22.8% versus 15.0%; for $\sigma = 1.4$ ($\rho \approx 0.63$), ML detected 25.2% versus 12.0%. Only at $\sigma = 1.8$ ($\rho \approx 0.20$)—a near-useless ranker—did the scheduled policy (18.9%) edge out ML (13.8%), which is the expected behaviour. The ML advantage is stable for rank correlations above roughly 0.40, well below what a model with AUC 0.735 would produce in practice. These results support the conclusion that the ML policy’s advantage is not sensitive to the specific noise assumption used.

6.7 Industry interpretation

To put the budget saving in practical terms, published subsea inspection costs run roughly US\$8 000–25 000 per kilometre. For a representative field of 1 000 km, a fixed-schedule campaign inspecting 15% of the network costs on the order of US\$2.3 million at the mid-range unit cost. Achieving the same detection with the ML policy requires inspecting only about 6% of the network, for a saving on the order of US\$1.3 million per campaign, or equivalently a markedly higher detection rate at unchanged spend. These figures are illustrative rather than exact,

but their scale is consistent with industry reports of AI-enabled AUV inspection programmes achieving substantial reductions in annual inspection spend. Table 4 summarises the illustrative cost translation.

Table 4: Illustrative per-campaign cost of reaching current-practice detection, for a representative 1 000 km subsea field. ML matches the scheduled-policy detection at roughly 60% lower budget.

Unit cost (US\$/km)	Scheduled campaign	ML campaign	Saving
8 000	US\$1.20 M	US\$0.48 M	US\$0.72 M
15 000	US\$2.25 M	US\$0.90 M	US\$1.35 M
25 000	US\$3.75 M	US\$1.50 M	US\$2.25 M

6.8 Discussion

For an integrity engineer, the practical message is that risk-prioritized routing extracts substantially more safety value from the same AUV inspection budget, and that the risk drivers are dominated by what the pipeline carries and how it is classified rather than by age alone. The framework should be applied where reliable segment attributes are available and where the objective is to rank many segments under a binding budget; it should not be treated as a replacement for detailed engineering assessment of an individual critical asset.

The framework produces an inspection allocation—a priority-ranked list of segments to inspect—rather than a feasible AUV route in the path-planning sense. [It is worth clarifying what the framework produces: a ranked list of segments to inspect—an inspection allocation—not a ready-to-fly AUV mission plan. Turning a ranked list into an actual route requires solving a path planning problem with travel distance, connectivity and vehicle endurance constraints. That routing step is a natural engineering extension of the current work, but is out of scope here. The framework’s claims are therefore limited to inspection prioritisation and budget allocation.](#) The main limitations—a representative rather than operator-specific network, US-based training data, a graph-traversal rather than physics-based AUV model, and a severity- rather than frequency-based risk target—each point to clear directions for follow-on work. [One notable absence is any feedback loop: as segments are inspected and anomalies are found or not found, those results could update the risk estimates through a Bayesian mechanism. A clean inspection result should lower a segment’s priority; a detected anomaly should raise it. Building that adaptive update into the system is identified as a valuable next step.](#)

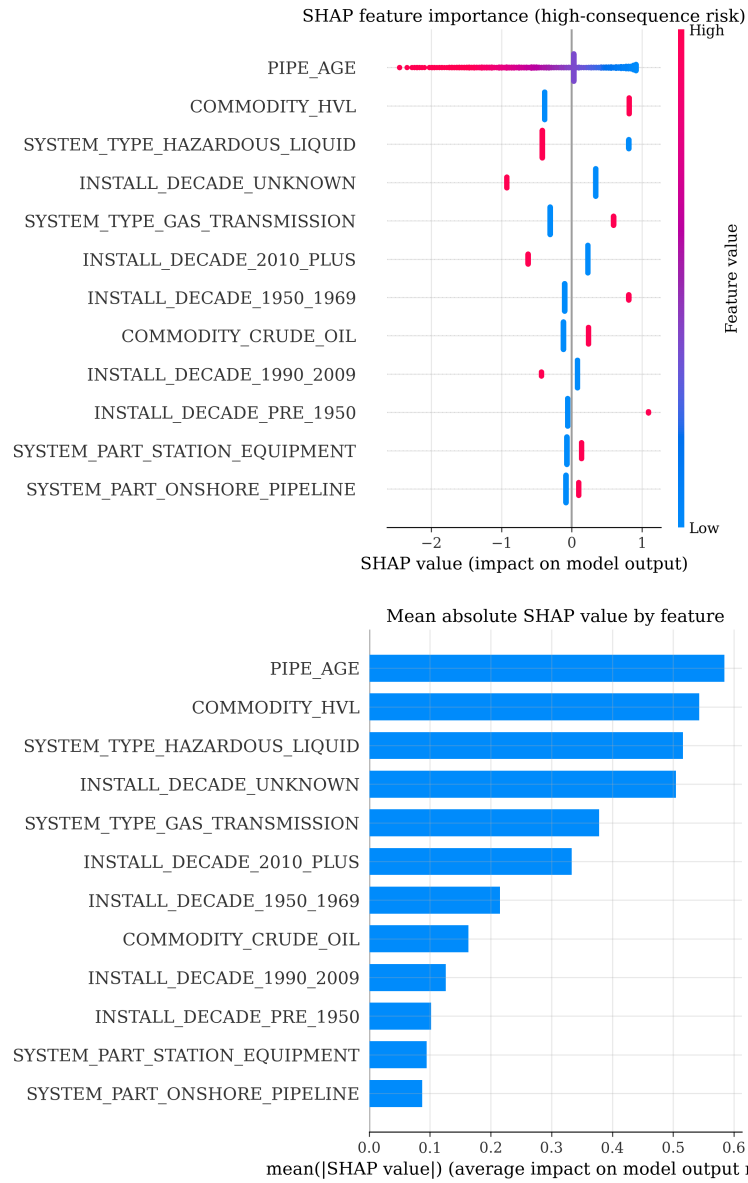


Figure 5: SHAP explanation of the model. Top: per-record feature attribution (beeswarm), where each point is one incident coloured by its feature value. Bottom: mean absolute SHAP value per feature, ranking overall influence on the predicted high-consequence risk.

Digital twin of the subsea network coloured by segment risk

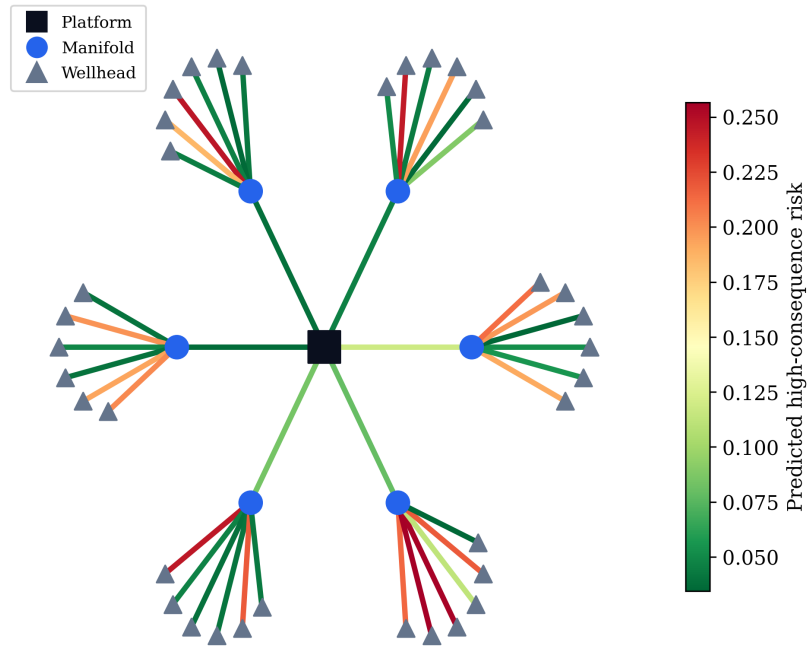


Figure 6: Digital twin of the representative subsea pipeline network. Nodes are the platform (square), well-cluster manifolds (circles) and wellheads (triangles); edges are pipeline segments coloured by predicted high-consequence risk.

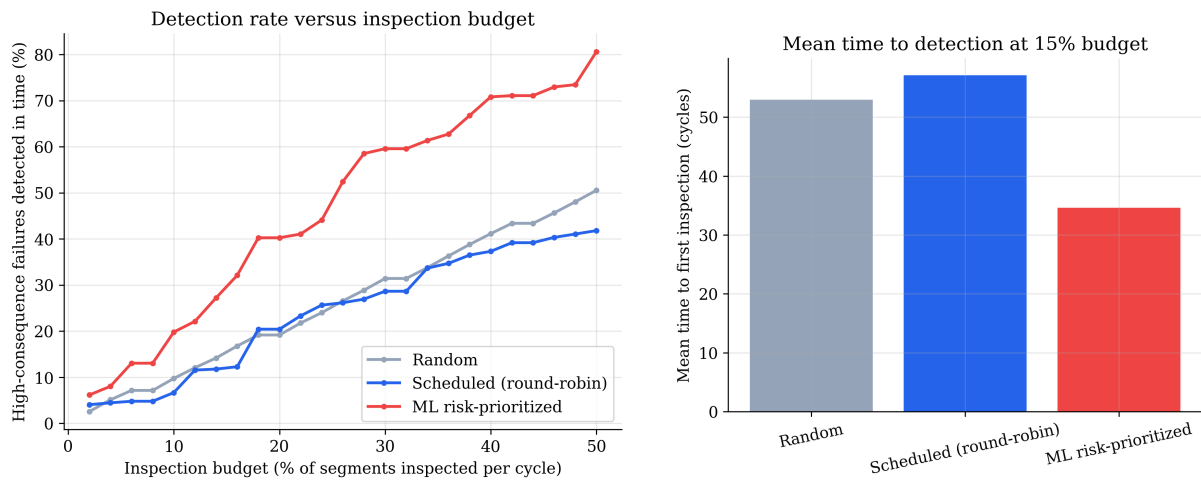


Figure 7: AUV inspection policy comparison. Left: high-consequence failures detected in time as a function of inspection budget. Right: mean time to detection at a 15% budget.

7 Conclusion and future work

This paper has presented a complete, open and laptop-deployable framework that links real pipeline incident data, interpretable machine learning and a graph-based digital twin to prioritize AUV inspection of subsea oil and gas pipelines. Three conclusions follow. First, a high-consequence failure model trained on 7813 public PHMSA records reaches a useful cross-validated ROC-AUC of about 0.74, with transported commodity and system type—not age alone—as the dominant, and explainable, risk drivers. Second, embedding these risk estimates in a digital twin and directing AUV inspection by risk approximately doubles the detection of high-consequence failures relative to fixed-schedule inspection at equal budget, or delivers current-practice detection at roughly 60% lower inspection effort. Third, because the entire system is built from free software and public data, this capability is reproducible and accessible to operators without commercial digital-twin budgets.

Three future directions are identified. The framework can be extended to use operator-specific data under appropriate data-sharing arrangements, and to incorporate exposure denominators so that absolute failure frequency, not only conditional severity, is modelled. The single inspection agent can be generalised to a coordinated team of AUVs and surface drones with explicit path and endurance constraints, converting the current inspection-allocation output into a full mission plan. Finally, the framework can be connected to a physically realised, additively manufactured inspection-vehicle platform, maintaining a deliberate link to manufacturing-engineering research and providing a natural foundation for a follow-on study. [The Python code, data-processing scripts and simulation configuration files are available from the corresponding author on request.](#)

References

- [1] El-Abbasy M S, Senouci A, Zayed T, Mirahadi F and Parvizsedghy L 2014 Artificial neural network models for predicting condition of offshore oil and gas pipelines *Autom. Constr.* **45** 50–65
- [2] Senouci A, Elabbasy M, Elwakil E, Abdrabou B and Zayed T 2014 A model for predicting failure of oil pipelines *Struct. Infrastruct. Eng.* **10** 375–387
- [3] Tao F, Zhang H, Liu A and Nee A Y C 2019 Digital twin in industry: State-of-the-art *IEEE Trans. Ind. Inform.* **15** 2405–2415
- [4] Jacobi M 2015 Autonomous inspection of underwater structures *Robot. Auton. Syst.* **67** 80–86
- [5] Pipeline and Hazardous Materials Safety Administration 2025 *Pipeline Incident Data: Distribution, Transmission & Gathering, LNG, and Liquid Accident and Incident Data* US Department of Transportation, <https://www.phmsa.dot.gov/data-and-statistics/pipeline/distribution-transmission-gathering-lng-and-liquid-accident-and-incident-data> (accessed 22 September 2026)
- [6] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: Synthetic minority over-sampling technique *J. Artif. Intell. Res.* **16** 321–357
- [7] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32
- [8] Chen T and Guestrin C 2016 XGBoost: A scalable tree boosting system *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* 785–794

- [9] Lundberg S M and Lee S-I 2017 A unified approach to interpreting model predictions *Adv. Neural Inf. Process. Syst.* **30** 4765–4774
- [10] Hagberg A A, Schult D A and Swart P J 2008 Exploring network structure, dynamics, and function using NetworkX *Proc. 7th Python in Science Conf. (SciPy2008)* 11–15
- [11] Pedregosa F *et al* 2011 Scikit-learn: Machine learning in Python *J. Mach. Learn. Res.* **12** 2825–2830