

APPLICATIONS OF MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE TO PIPELINE GEOHAZARD INTEGRITY MANAGEMENT PROGRAMS

Sarah Newton
Cambio Earth
Calgary, AB

Aron Zahradka
Cambio Earth
Victoria, BC

Corey Scheip
BGC Engineering
Nashville, TN

Andrew Johnson
Cambio Earth
Vancouver, BC

Doug Dewar
Pembina Pipeline Corp
Calgary, AB

ABSTRACT

Pipeline geohazard integrity management programs increasingly rely on monitoring and remote sensing data to detect hazards affecting vast pipeline networks. While these datasets provide unprecedented visibility into ground movement and hydrotechnical processes, many programs are now constrained by the ability to consistently interpret data at scale. Machine learning (ML) and artificial intelligence (AI) can help automate interpretation while keeping human oversight in integrity management workflows.

This paper reviews emerging ML/AI applications relevant to pipeline geohazard management and situates them within the program lifecycle. Three representative case studies are presented: (1) a deep learning model that screens in-line inspection inertial measurement unit bending strain data for signatures consistent with ground movement; (2) a probabilistic ML model that predicts annualized probability of failure for pipeline watercourse crossings; and (3) a computer vision model that automates mapping of active landslide movement from lidar change detection.

These examples illustrate how ML enables pattern recognition across large datasets while highlighting implementation requirements including high-quality training data, explicit treatment of uncertainty, clear domain of application, and integration into existing risk management workflows. Lessons learned from these applications provide a roadmap for responsible scaling of ML/AI in pipeline geohazard programs.

Keywords: Geohazards, Machine Learning, Artificial Intelligence, Pipeline Integrity, Probability of Failure, Risk

1. INTRODUCTION

Many pipeline operators have implemented risk-informed geohazard management programs (GMPs) that combine systematic hazard inventories with screening-level assessments of probability of failure (PoF) and consequence to prioritize

monitoring, inspection, and mitigation effort. Experience applying screening PoF algorithms across large inventories has demonstrated that these frameworks can support triage of thousands of geohazard sites and help focus limited engineering resources on the highest risk sites while reducing unnecessary effort at benign sites [1,2].

Despite progress, pipeline geohazard failures continue to occur. Recent operational datasets indicate that extremely to very slow-moving [3] landslides remain the dominant cause of geohazard-related pipeline failures in North America, while failures from hydrotechnical hazards have declined over the past decade [4–6]. This shift implies that mature programs must increasingly manage long-duration, low-visibility ground movements that accumulate strain over years to decades and can be difficult to detect through visual inspections alone.

At the same time, the volume of monitoring and remote sensing data available to integrity teams has expanded dramatically. The use of multi-vintage lidar, InSAR, higher-resolution optical imagery, and in-line inspection (ILI) inertial measurement unit (IMU) tool runs has shifted many programs from being data-limited to interpretation-limited: the primary constraint is no longer the availability of observations, but the ability to screen, interpret, and act on those observations consistently and at scale [4].

Interpretation-limited programs face several practical challenges: screening high volumes of data to identify the small fraction of signals that indicate credible pipeline integrity concerns; maintaining consistency across analysts, regions, and changing personnel; and allocating constrained assessment and mitigation budgets where opportunities for risk reduction are highest.

This paper's central thesis is that select machine learning (ML) and artificial intelligence (AI) applications can accelerate the maturity of pipeline GMPs by removing interpretation bottlenecks associated with high-volume monitoring datasets. ML/AI enables scalable pattern recognition across large datasets,

allowing integrity teams to prioritize engineering review and focus resources where intervention is most likely to reduce risk.

Importantly, the value of ML/AI in pipeline integrity is not realized through automation of engineering judgment, but through improved triage and earlier identification of potential hazards within established risk management workflows. When deployed within clearly defined domains of application, supported by transparent validation, and combined with human review and accountability, ML/AI provides a practical mechanism to transition geohazard programs from reactive hazard identification towards primarily proactive monitoring and mitigation.

Pipeline GMPs tend to evolve through recognizable stages of maturity. Early programs focus primarily on hazard discovery and event-driven response, while more mature programs develop systematic inventories, risk-informed prioritization frameworks, and integrated monitoring workflows. As remote sensing and monitoring technologies have expanded, many programs have reached a stage where the limiting factor is no longer data acquisition but the ability to interpret large volumes of monitoring data consistently and at scale. The following sections introduce a maturity model for pipeline GMPs and illustrate how ML/AI applications can accelerate transitions between these stages.

2. PIPELINE GMP MATURITY AND ML/AI

Pipeline GMPs tend to progress through recognizable stages of maturity, although operators and asset systems may be at different stages at any given time. In this paper, program maturity is defined by the degree to which hazards are systematically identified, risk is quantified and used to allocate resources, monitoring is integrated to detect change, and decisions are executed through repeatable workflows rather than ad hoc responses. Mature programs progressively shift effort away from reactive hazard discovery and toward continuous monitoring, proactive risk control and mitigation.

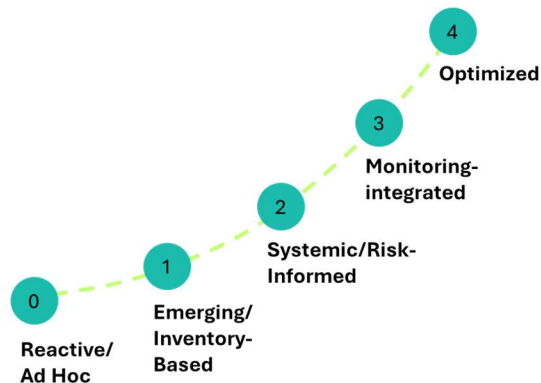


FIGURE 1: PIPELINE GMP MATURITY MODEL.

Figure 1 summarizes a practical maturity model representing this progression and highlights where ML/AI capabilities can accelerate transitions between maturity states. In practice, ML/AI does not replace the analytical frameworks used

within GMPs; rather, it accelerates progression between maturity levels by scaling interpretation of high-volume datasets and improving consistency in screening and monitoring workflows.

Maturity model summary (Levels 0–4)

0 – Reactive/Ad Hoc: Geohazard management is primarily event or failure-driven. Hazard inventories are incomplete or absent, and response actions are typically triggered by visible damage, third-party reports, or emergency situations. Decision making relies heavily on local knowledge and individual experience. ML/AI applications are generally not applied at this stage because structured datasets and labelled examples required for model development are limited.

1 – Emerging/Inventory-based: Operators begin systematic identification of geohazards such as landslides and watercourse crossings using desktop mapping, historical records, and field reconnaissance. Screening and prioritization are largely qualitative and rely on subject matter expert (SME) interpretation. ML/AI can support this stage by accelerating corridor-scale hazard inventory development and improving consistency in initial screening.

2 – Systemic/Risk-informed: Screening-level PoF or risk assessment frameworks are implemented to prioritize inspection, monitoring, and mitigation resources across large site inventories. Site records are systematically documented, and decision criteria become repeatable across regions. ML/AI can accelerate this stage by generating rapid screening estimates and reducing inter-analyst variability in hazard characterization.

3 – Monitoring-integrated: Remote sensing and monitoring datasets such as IMU, InSAR, lidar change detection (LCD), and geotechnical instrumentation are integrated into the GMP to find ground movements and trigger follow-up investigation. Trigger-action response plans define how monitoring results influence engineering decisions. ML/AI applications automate initial interpretation of high-volume monitoring datasets, allowing engineers to focus on likely integrity threats.

4 – Optimized: Monitoring and screening processes operate as a continuous, integrated workflow. High-volume datasets are automatically triaged using ML models that prioritize anomalies, update risk assessments, and route results to defined engineering decision gates. Human SMEs remain responsible for final decisions, but ML/AI reduces screening and interpretation bottlenecks and enables earlier and more consistent risk control and mitigation actions across large pipeline systems.

3. LANDSCAPE OF ML/AI SOLUTIONS

ML/AI encompass a broad set of computational approaches designed to identify patterns within large datasets. In pipeline GMPs, these approaches are most useful when applied to repeatable interpretation tasks involving large monitoring or remote sensing datasets.

ML is a sub-field of AI, and because most of today's active AI research falls within ML, they are often used interchangeably in practice [7]. Different ML architectures are suited to different data types commonly encountered in pipeline monitoring. Tree-based models are well suited to structured tabular datasets such

as hydrologic, watershed, or pipeline attribute data. Deep learning approaches such as convolutional neural networks (CNNs) or transformer architectures are particularly effective for spatial pattern recognition tasks involving imagery or gridded terrain data, including LCD products. Transformers and sequential architectures such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks can be used to analyze time-series datasets such as instrumentation monitoring records or weather data.

In practice, however, engineers interact with these technologies primarily through the operational tasks they support rather than through the underlying model architecture. Within GMP workflows, ML/AI applications typically support several recurring functional tasks.

Hazard identification and inventory development: ML models can assist in detecting and delineating potential geohazards across large geographic areas using remote sensing datasets. Examples include landslide susceptibility models developed by geological surveys, NASA's Semi-Automated Landslide Detection (SALaD), and automated land surface disturbance detection products such as the OPERA Land Surface Disturbance dataset. ML has also been applied in the US Appalachian region by Cambio Earth and BGC Engineering to LCD datasets to automatically map areas of active landslide movement and accelerate corridor-scale hazard inventories.

Screening and prioritization: ML can synthesize watershed, stream, and pipeline characteristics to predict screening-level PoF for pipeline hydrotechnical hazards [8]. These models allow operators to rapidly prioritize large inventories of hydrotechnical sites for follow-up actions when field inspection data are not yet available.

Monitoring and anomaly detection: Many GMP datasets are collected repeatedly through time, including lidar surveys, satellite imagery, and ILI data. ML models can screen these datasets to identify anomalies or changes that may indicate evolving geohazard conditions. For example, models can identify strain signatures in ILI IMU data that may indicate ground movement interacting with a pipeline [9].

Forecasting and early warning: Hybrid modelling approaches combining physical process models with ML components are increasingly used to forecast environmental hazards such as floods and landslides. For example, the National Oceanic and Atmospheric Administration's National Water Model uses data-driven components to support large-scale flood forecasting, while NASA's Landslide Hazard Assessment for Situational Awareness (LHASA) model produces global landslide nowcasts and short-term forecasts.

Knowledge management and decision support: A different class of AI tools, generative AI and large language models (LLMs), addresses the challenge of synthesizing large volumes of unstructured information. Decades of pipeline operation generate extensive textual records including inspection reports, geotechnical investigations, mitigation designs, and regulatory documentation. LLM-based systems can extract structured information from these documents, summarize historical site

records, and support rapid retrieval of relevant information for engineering review.

Unlike predictive ML models, which identify patterns in numerical or spatial datasets, LLMs operate primarily on text and other unstructured data. Their role within GMPs is therefore typically limited to information management and decision support rather than quantitative hazard prediction.

The practical value of ML/AI within pipeline integrity programs ultimately depends on integration with established GMP workflows. Model outputs must connect directly to hazard characterization, monitoring decisions, or mitigation planning processes. When deployed in this manner, ML primarily functions as a triage layer that accelerates interpretation of large datasets while preserving human accountability for engineering decisions.

4. CASE STUDIES

Three case studies of ML used to accelerate GMP maturity are described in the following subsections.

4.1 Landslide Impact Detection in IMU

Extremely slow to very slow-moving landslides are a leading cause of geohazard-related pipeline failures worldwide [4,10,11]. Landslides often threaten pipeline integrity through progressive strain accumulation over years to decades. Because ground movements can be subtle and surface expression can be limited, confirming landslide activity and its interaction with a buried pipeline often requires pipeline monitoring rather than visual inspections alone [4,12,13].

ILI IMU tools allow the measurement of pipeline geometry changes and subsequent bending strain signatures associated with external loading. IMU has become a critical tool for evaluating landslide activity and interactions because it can be applied along entire pipeline alignments. In geotechnical terms, the pipeline becomes a near horizontal slope inclinometer.

However, IMU assessment is subject to interpretation and typically relies on industry SMEs. A single tool run typically generates data every 0.02 to 0.2 m for 10s to 100s of kilometers. Following collection, raw IMU data is processed by vendors to produce tabular data documenting the pitch and azimuth of the pipeline. From pitch and azimuth, the curvature of the pipeline is calculated using a summed gage length. The curvature is referred to as a "strain" and is separated into horizontal, vertical and total strains. In the authors' experience, less than half of IMU ILI runs undergo bending strain or run-to-run movement analysis. Even when a bending strain report is commissioned by the pipeline operator, few of these reports will be analyzed further by geohazard SMEs. The vendor bending strain reports generate many bending strain features, the vast majority of which are associated with benign construction-related features such as roped overbends, road bore tie-ins or deformed field bends that did not fit the ditch bottom profile. Conventional screening approaches use simple strain thresholds to flag features for review [14], which produce high false positive rates (e.g., ~5% precision, requiring engineers to review ~20 non-landslide features for each landslide feature) [9] and create a

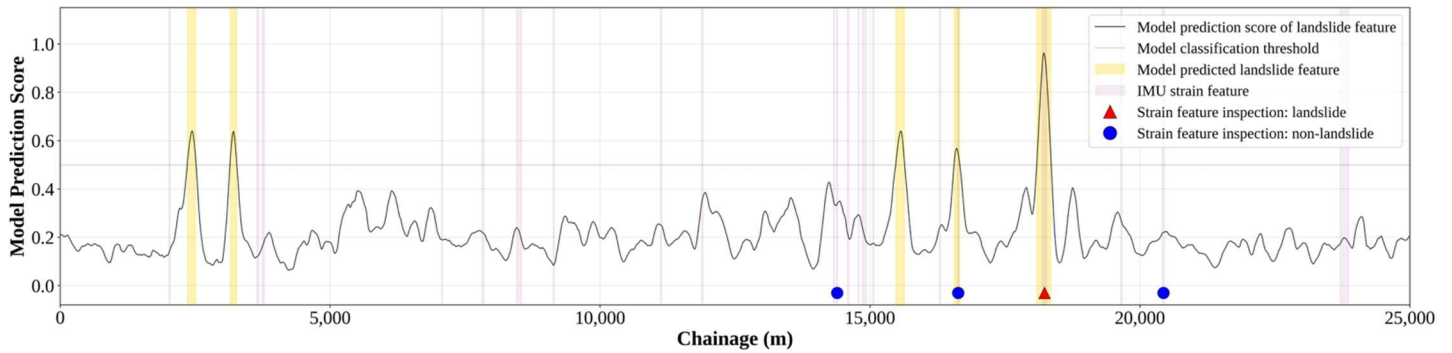


FIGURE 2: ILLUSTRATION OF THE DEEP LEARNING MODEL RESULTS PREDICTING LANDSLIDE IMPACT LIKELIHOOD (BLACK LINE) ACROSS A REPRESENTATIVE 25 KM SECTION OF IMU.

bottleneck for manual SME review.

To address this interpretation bottleneck, a deep learning model was developed to automatically screen IMU bending strain features for signatures consistent with ground movement, as illustrated in Figure 2 [9]. The model uses a CNN trained on 30,000 labelled bending strain features previously classified in Cambio. Cambio is a geohazard management platform used by 40 pipeline operators to support GMPs on more than 500,000 km of pipelines. The training set included 478 examples of features classified by SMEs as being related to landslide impacts. Training and testing used a 25% spatially distinct hold-out split of the labelled dataset. On the withheld test data, the model achieved approximately 90% recall, 95% specificity, 95% overall accuracy and an ROC-AUC of 0.96 for identifying landslide-impacted strain features. Further details on the model architecture, training data, and validation results are provided in Zahradka et al. [9] and work on this model continues to advance [15].

Operationally, the primary value of an approach like this is the reduction of false positives while maintaining high recall for impacted features. The model reduces the number of strain features requiring SME review by approximately a factor of five compared to traditional screening criteria, enabling SMEs to focus on the most probable landslide and settlement-related integrity concerns.

Integration into the GMP workflow consists of extracting IMU bending strain features from ILI runs; the ML model then assigns a probability of landslide association enabling SMEs to review and document their interpretations for a reduced set of features. The findings update the site's hazard characterization and response level (for example, via operator-specific matrices aligned with API RP-1187 concepts). Where appropriate, follow-up actions are initiated, such as field inspections, additional monitoring, detailed assessments, or in extreme cases, stress relief and mitigation.

When compared to a GMP using traditional screening criteria, the ML model's false positive reduction is estimated to result in as many as 50 fewer analyst hours of IMU strain inspection effort for an average ILI run of 110 km.

Going forward, ILI IMU ML/AI analysis can be expanded to other ILI wall anomaly feature interactions including but not limited to dents, wrinkles, buckles, ovality, metal loss and

cracking (axial/circumferential). Additionally, stress/strain tool results could be incorporated into any ML/AI analysis.

4.2 Landslide Activity Mapping in Lidar Change Detection

Multi-vintage airborne lidar has become a cornerstone dataset for pipeline geohazard programs because it enables high-resolution topographic mapping and, when repeated through time, measurement of surface elevation change. LCD can reveal subtle landslide deformation (down to 0.1 m movements). Lidar and LCD have the distinct advantage of being able to support landslide identification in forested terrain where ground deformations are obscured by vegetation in satellite imagery [16,17].

LCD interpretation is commonly performed through manual review by landslide SMEs who delineate areas of active movement based on the presence of landslide-like features including scarps, hummocky/broken/uneven ground, toe bulges and zones of deposition. When comparing multiple datasets, patterns of elevation loss and gain consistent with landslide movement can be used. While SME interpretation is effective, it is time-consuming at corridor scale and is inconsistent across analysts.

A supervised computer vision ML model was developed to automate mapping of active landslide movement zones from LCD outputs. The model is trained on SME-classified and segmented movement polygons (movement zones rather than full landslide bodies), created by landslide specialists using LCD and supporting geomorphic context. The primary challenge for the model is distinguishing between landslide processes versus other sources of natural (e.g., erosion, subsidence) and anthropogenic terrain movements; this requires the model to learn the topographic change signatures typically associated with active deformation and the terrain morphology patterns visible in the lidar. The model produces segmented polygons indicating candidate movement areas suitable for analyst review and integration into GMP inventories.

The primary operational benefits are reduced manual interpretation workload for recurring LCD campaigns, improved consistency of movement mapping, and improved detection of small or subtle movements compared to manual screening. The

model provides a more complete basis for SME assessment and follow-up monitoring.

Integration into GMP workflows involves generation of LCD rasters from two lidar datasets of differing vintages, which the ML uses to produce landslide movement extents as geospatially referenced polygons. The landslide movement areas can then be reviewed by SMEs along with the LCD dataset as part of accelerating geohazard site inventory creation or supporting monitoring and remote/desktop inspection programs. Integration of these datasets directly into the geohazard management system used by the GMP yields further efficiencies.

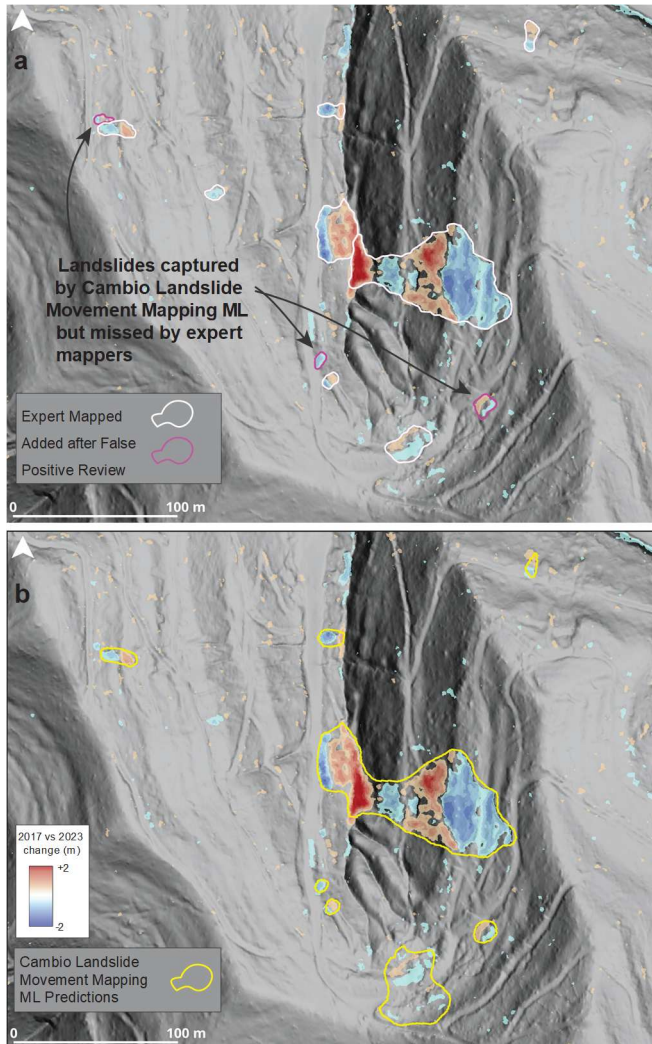


FIGURE 3: COMPARISON OF HUMAN AND ML MAPPED LANDSLIDE MOVEMENT AREAS FROM LCD.

After initial model training, the model detected 87% of manually mapped landslides. The data scientists employed an active learning loop, which consisted of human review of all false positives produced by the model ($n=3,100$). Upon review, over 1,800 of the false positives were found to be real landslides that the model had detected but the human mappers had missed (Figure 3). Adding these landslides to the model training data

resulted in a final detection rate of 92.5% of landslides. This result highlights a known limitation of large manual mapping exercises: mapper fatigue and variability between analysts. Arguably, the most complete inventory is produced by hybrid ML followed by SME review.

As lidar becomes more widely available (e.g., upcoming 5-year cadence of U.S. Geological Survey Next Gen 3DEP program) or less expensive to acquire, LCD will similarly become more accessible. Coupled with SME reviews, leveraging ML to quickly and consistently map landslides at scale will continue to improve GMP workflows.

4.3 Watercourse Crossing Hazard Screening

Hydrotechnical geohazards at pipeline watercourse crossings can compromise pipeline integrity through flood-induced scour that exposes or damages the pipeline. Geohazard prioritization at watercourse crossings is commonly based on the annualized PoF, defined as the product of the probability of pipeline exposure (PoE) in the watercourse and the vulnerability (V) of the pipeline to failure given exposure. PoE is estimated by computing scour depths for a range of flood return periods and comparing these to the pipeline depth of cover (DoC), while vulnerability is estimated from pipeline and loading characteristics.

Conventionally, estimating PoF requires field inspections to characterize channel geometry, measure DoC, and collect observations used to calculate scour depths and vulnerability. However, advances in bare-earth lidar and high-resolution satellite imagery [18] have greatly increased the number of identified watercourse crossings, averaging approximately one crossing every three kilometres [1]. The cost of field inspections for all identified crossings is disproportionate to the overall risk posed, particularly given that Newton et al. [2] found that only 5% of hydrotechnical sites account for approximately 93% of total system-wide PoF. The remaining 95% of sites collectively contribute only 7% of the total PoF, forming a "long tail" that presents a significant screening challenge in the early stages of geohazard programs when most sites lack field-inspected PoF estimates. Generating screening-level PoF estimates rapidly and efficiently is critical to ensure that geohazard management resources are deployed to the small number of sites most likely to have the highest PoF.

A probabilistic ML model was developed to predict PoF at watercourse crossings using widely available desktop-level data, reducing the need for field inspections to prioritize sites for detailed assessment. The model was trained on data from approximately 20,000 field inspections conducted across North America over the past decade, systematically documented using a standardized methodology and reviewed by engineers and geoscientists using Cambio [8].

Input features were drawn exclusively from datasets available at scale without field inspections. The features were organized into three categories: (1) pipeline characteristics at the crossing (e.g., installation date, diameter, wall thickness, and material), (2) watercourse characteristics (e.g., geographic location, stream order, stream gradient, and flood quantiles from

screening-level automated flood frequency analysis), and (3) watershed characteristics derived from climatic, land cover, soil hydrological, and topographic datasets along with catchment area.

To enhance interpretability and align predictions with established physical process mechanics, separate ML models were developed to predict physically meaningful components of PoF: minimum DoC, the scour depth curve (scour depth as a function of flood return period), and pipeline vulnerability. This structure preserves physical interpretability while enabling rapid screening. These component predictions are integrated following the PoF framework described above to produce the final PoF estimate. This multi-model structure enables the modelling chain to independently capture the physical processes underlying each component while maintaining transparency in how the final PoF is derived.

In high-consequence engineering applications, understanding and quantifying the uncertainty of model predictions is essential to evaluate their reliability [10,11]. Probabilistic ML was therefore adopted to predict full probability distributions rather than single-point estimates, using quantile regression with XGBoost [19–21]. Predicting distributions captures the inherent uncertainty in the predictions and enables adjustment of conservatism to suit the risk tolerance of the application. For screening purposes, the 70th percentile of each site's predicted PoF distribution was selected to introduce a controlled conservative bias, prioritizing the identification of potentially high-risk sites for follow-up field inspection.

Model testing was performed on a withheld test dataset of 4,048 sites (20% of the available data). The predicted PoF distributions demonstrated good calibration against field-inspected results, with the overall distribution of predicted PoF aligning well with the actual distribution. The predicted PoF values were binned into hazard categories based on order-of-magnitude ranges (e.g., Category 1: $\text{PoF} > 10^{-2}$, Category 6: $\text{PoF} < 10^{-6}$). The matrix of predicted versus field-assessed hazard categories showed that the model effectively identifies the highest-risk sites while maintaining a slight conservative bias, consistent with the selection of the 70th percentile. Further details on the model architecture, training data, and validation results are provided in Zahradka et al. [8].

In the case study of a pipeline system presented in Zahradka et al. [8], 1,603 previously identified watercourse crossings were screened and prioritized using the model; field inspection costs were reduced by 84%, saving an estimated \$2M CAD in field inspections of very-low risk sites.

5. DISCUSSION

5.1 ML/AI as a Natural Progression in GMP Workflows

The three case studies presented here illustrate a common pattern: ML/AI creates value in pipeline GMPs when it removes interpretation bottlenecks in repeatable, high-volume data streams and routes engineering attention toward the subset of signals that are most likely to require action. Applications as presented here tend to reduce subjectivity and increase

repeatability – key components of building a defensible GMP. In this sense, ML/AI is best understood as an accelerator, enabling programs to transition from spending resources on inventory and screening toward monitoring-integrated, proactive risk management workflows. Viewed through the maturity model introduced earlier, the primary contribution of ML/AI is to accelerate transitions between stages by automating interpretation tasks that would otherwise constrain monitoring-integrated programs.

5.2 Implementation Best Practices

Successful deployment of ML/AI within a GMP depends largely on the quality of the ML implementation itself, and this paper's case studies offer practical examples of implementation.

Model architecture may be inherently complex and opaque, but the process used to develop, train, validate, and scope the model should be transparent [22]. Operators evaluating ML/AI solutions, whether developed internally or by third-party vendors, should be able to clearly understand:

- The data on which the model was trained and how it was curated,
- The confidence or uncertainty associated with individual predictions,
- How the model was tested, including whether evaluation was performed on data used for testing only (and not seen by the model during training),
- The metrics used to assess performance and why those metrics are relevant to the intended application, and
- The domain of application within which the model is expected to perform reliably.

The following subsections discuss these considerations in detail, drawing on the case studies to illustrate how each was addressed in practice.

5.2.1 Training data quality

The performance ceiling of any ML model is set by its training data. Across the case studies presented in this paper, the most consequential implementation decisions were not about model architecture or hyperparameter tuning, but about how training datasets were assembled, labelled, and scoped. Domain expertise and specificity are critical to this process.

Both training data diversity and volume are influential for model performance. Diversity and representativeness of training examples must match the conditions the model will encounter in deployment. The watercourse PoF model was trained on approximately 20,000 field inspections of unique sites; there were many additional repeat visits recording changed conditions at a site over time. While repeat inspections add some information, a new inspection at a site with materially different geomorphic, hydrologic, and pipeline characteristics contributes far more to model generalization and is the metric that matters. Practitioners assembling training datasets should prioritize breadth of conditions over raw sample count. However, it is paramount to recognize the subjective nature of human inspections or analyses. These data are inherently subjective and match that SME's particular set of experiences and training.

Therefore, there is inherent scatter, or noise, in all training datasets. This is acceptable in ML/AI applications if sufficient volumes of training data are provided to the model.

Part of model generalization is learning the variety of interpretations for a given problem. For example, the active landslide mapping model was shown manually drawn landslide boundaries, which are highly subjective and vary greatly across different mappers. To account for this, the model was shown over 6,400 unique landslide extents mapped by a few different specialists, capturing at least some of the diversity of individual biases.

GMPs are primarily concerned with relatively rare events (e.g., a landslide impacting a pipeline). In classification or detection applications, this results in a persistent challenge of high class imbalances between positive and negative samples. The IMU landslide detection model, for example, had 30,000 bending strain features screened or inspected by specialists but only 478 confirmed landslide-impact examples (positive samples) – a class ratio of roughly 60:1. At this imbalance, the quality and consistency of the positive samples are very important, and a small number of mislabelled positives can meaningfully affect model quality. With a 25% hold-out split, approximately 120 positive samples remain for testing, which is sufficient to coarsely estimate recall in aggregate but limits the ability to characterize performance across different geographic regions beyond Appalachia. This tension between reserving data for robust testing and retaining it for training is inherent to rare event geohazard problems and should be acknowledged rather than ignored.

Data governance and provenance underpin the above considerations. The watercourse PoF model benefits from a training dataset collected over two decades using a standardized inspection methodology and reviewed by senior engineers and geoscientists using Cambio. ML readiness begins well before model development with consistent data collection and organization protocols, documentation of how inspection practices and underlying empirical models have evolved, and structured storage of all data in a software system backed by a relational database that enforces data integrity. Where labelling standards or analytical methods have changed over the period of record, care must be taken to reconcile historical data, as was done for the PoF model by recalculating vulnerability using the latest empirical model for consistency across decades of inspections.

5.2.2 Consideration of uncertainty

In high-consequence engineering applications, a single-point prediction is often insufficient for defensible decision-making [22]. Pipeline geohazard risk management requires not only an estimate of the hazard or PoF, but also an understanding of how much confidence can be placed in that estimate. This distinction separates ML applications in pipeline integrity from most commercial ML use cases, where the cost of an incorrect prediction might be low and a best estimate is adequate.

Many ML classification models, such as the active landslide mapping model (Section 4.2), output a score between 0 and 1

that is often interpreted as a probability. However, these scores are discriminative – they are optimized to rank and separate classes, not to represent calibrated likelihoods. A model score of 0.3 does not necessarily mean there is a 30% chance the feature is landslide-related. For screening applications where the primary objective is to rank features for review and select a classification threshold, this distinction may be acceptable provided the threshold is calibrated against known outcomes and the model's discrimination performance is validated on independent test data. Where model outputs are used more directly in quantitative risk calculations, however, calibrated probabilistic predictions become important.

The watercourse PoF model (Section 4.3) was designed from the outset as a probabilistic model, using quantile regression [19] to predict full probability distributions for each component of PoF rather than single-point estimates. Predicting distributions enables operators to adjust conservatism to match their risk tolerance. For screening, the 70th percentile of each site's predicted PoF distribution was usually selected to introduce a controlled conservative bias, prioritizing identification of potentially high risk sites at the cost of some additional false positives. A different percentile could be selected for a different application or risk tolerance.

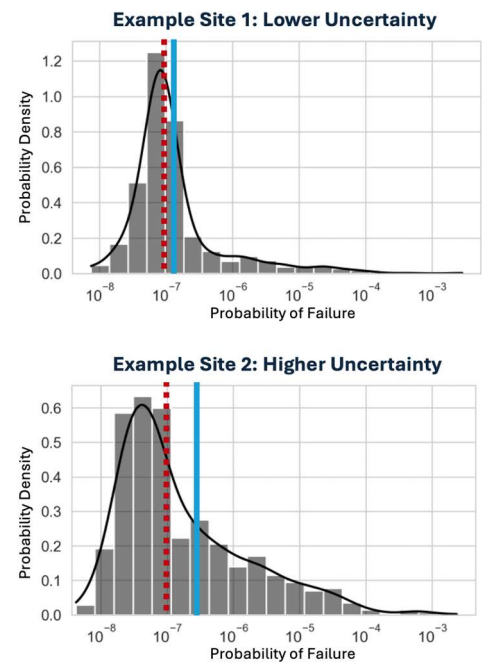


FIGURE 4: PREDICTED POF DISTRIBUTIONS FOR TWO EXAMPLE SITES. MEDIAN PREDICTIONS ARE SHOWN AS RED DASHED LINES AND THE 70TH PERCENTILES ARE SHOWN AS SOLID BLUE LINES.

Because the PoF model chains three component sub-models (DoC, scour depth, and vulnerability), uncertainty from each component propagates and compounds in the final PoF estimate. The probabilistic framework handles this through Monte Carlo sampling from component distributions, which also means that

the final PoF uncertainty reflects accumulated uncertainty across all stages. Validation of the predicted distributions against field-inspected results demonstrated good calibration [8], which is a distinct and arguably more meaningful form of validation than conventional performance metrics alone: it indicates that the predicted uncertainty is itself trustworthy, not just that the central tendency is close.

Figure 4 illustrates why considering the full prediction distribution matters for decision-making. Two real example sites have similar median predicted PoF values (a common output of non-probabilistic ML models) but markedly different prediction uncertainty. The narrow-distribution site can be triaged with relatively high confidence, while the wide-distribution site – where the predicted PoF could plausibly span multiple hazard categories – may warrant field verification or more conservative treatment. Relying on the median alone would treat these two sites identically, obscuring a meaningful difference in the confidence of the prediction and potentially leading to under- or over-allocation of engineering resources.

5.2.3 Model domain

Section 5.2.1 introduced the importance of assembling training data that is representative of the conditions a model will encounter in deployment. ML models generally interpolate well within the conditions represented in their training data but extrapolate poorly beyond them. Operators need to understand where the model has been tested, and where it has not, so they can assess whether their assets and conditions fall within the tested envelope. Additionally, GMPs operate over decades, during which precipitation patterns and flood frequencies shift, land use evolves, and pipeline conditions degrade. Temporal validity should be treated as a dimension of model domain that requires ongoing monitoring and periodic revalidation or retraining.

The watercourse PoF model's training dataset (Section 4.3) spans a broad range of pipeline characteristics, watercourse types, and physiographic settings across North America. The geographic and pipeline attribute coverage was explicitly characterized so that operators can evaluate whether a given pipeline system falls within the domain of tested performance. The IMU landslide detection model was evaluated with regional geographies assessed separately, and the characteristics of the pipelines in the training and test datasets were documented. The initial model was developed predominantly on large-diameter transmission pipelines in Appalachia [9] and recognition of this limitation has accelerated deliberate collaborations with other operators to expand the training dataset and build broader applicability of the model. The LCD model was trained and tested in a targeted region in Appalachia, primarily on relatively small earth slides. Efforts are ongoing to expand the model into other geographies and landslide types, including large, very slow-moving translational earth slide complexes in Western Canada and debris flows in North Carolina, which will be accompanied by updated performance metrics for the added regions.

5.2.4 Test set independence

Beyond defining where a model applies, the credibility of reported performance depends on how the test dataset itself is constructed. Ensuring that the test dataset is truly independent of the training data is, in the authors' experience, the most common failure point in applied ML projects for geohazard applications. The consequence is a model that appears to perform well during development but fails to deliver that performance when deployed in a real-world scenario where data from the test environment was not available during training. This failure mode is common in large part because random splitting of data into training and test partitions is the default behaviour in virtually all ML frameworks and introductory resources; mitigation comes from integrating both domain expertise and experienced ML practitioners into projects.

The fundamental question to ask when designing a train-test split is: if I know something about a training example, does it tell me anything about a test example? If the answer is yes, there is information leakage and the test evaluation will overestimate real-world performance. The IMU landslide detection model illustrates why random splitting is insufficient when data have spatial or temporal structure. Repeat IMU tool runs through the same location on the same pipeline produce bending strain features that are near-duplicates; if some runs appear in training and others in testing, the model can trivially memorize location-specific patterns rather than learning generalizable signatures of ground movement. Parallel pipelines traversing the same landslide introduce a subtler form of leakage: if a large landslide is interacting with two similar pipelines, features from both are not independent even though they come from different assets. The appropriate strategy in this case was a careful spatial split, ensuring that all features from a given geographic area were assigned entirely to training or testing. An important further test, applied subsequently, was evaluation on a full pipeline system never seen by the model to confirm generalization to new assets.

5.2.5 Performance metric selection

The choice of how performance is measured can determine whether evaluation results are meaningful or misleading. There is no single metric that universally captures model performance. The metric must be genuinely reflective of the real-world value the model is expected to generate and should be defined in collaboration with domain experts and stakeholders, ideally at the onset of a project before model development begins.

"Accuracy" is nearly synonymous with "model performance" in common usage, but it is almost never an appropriate primary metric in a GMP context. When a dataset is imbalanced, or when one type of model error is more costly than another (both of which are nearly always true in real-world geohazard applications), accuracy can be extremely misleading. As an example, the IMU landslide detection model attained a seemingly impressive overall accuracy of 95.1% on its test dataset. However, due to the heavily imbalanced nature of the test data (112 landslide examples among 7,024 non-landslide features), a naive model that only ever predicted "non-landslide" would have achieved 98.4% accuracy despite being entirely

useless and potentially dangerous if deployed. Metrics that separately characterize detection of the minority class, such as recall and precision, along with confusion matrices segmented by class, provide a more informative picture of how the model behaves on the cases that matter most.

The LCD model illustrates a different dimension of metric design. A key evaluation question is how many of the hand-mapped landslides the model detects, but not all landslides are equally consequential for pipeline integrity. Missing a large, deep-seated landslide intersecting and interacting with a pipeline to failure is far worse than missing a small, shallow surficial movement. Recall was therefore evaluated as a curve across the range of landslide areas, weighting the evaluation toward the features that matter more (e.g., larger landslides) for pipeline integrity decisions.

5.2.6 Practical model testing considerations

Beyond statistical baselines, any ML model deployed in a GMP should be benchmarked against the existing screening process it aims to replace or augment, not evaluated in isolation. This comparison is what gives operators confidence that the model adds genuine value over current practice and provides the basis for evaluating the trade-off between model complexity, cost, and screening quality. For the IMU model, the choice to benchmark against random chance would have made the model seem very impressive but would have been largely meaningless; the industry's existing threshold-based screening criteria performed far better than chance and are a far more appropriate comparison. The ML model's contribution must be demonstrated relative to that incumbent process in terms of both detection performance and the reduction in false positives that drive workload for engineering review.

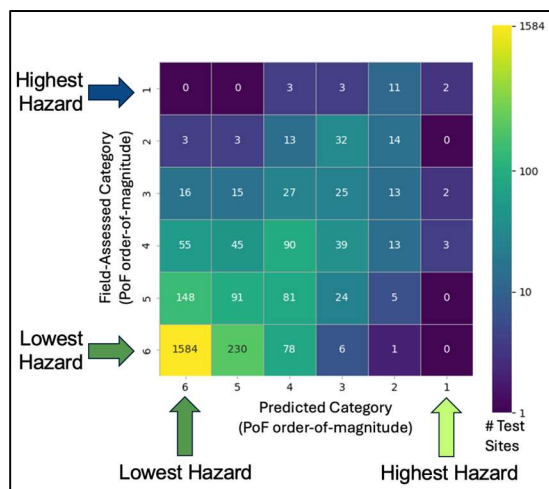


FIGURE 5: TEST RESULTS OF THE WATERCOURSE POF MODEL PRESENTED AS A MATRIX OF THE NUMBER OF SITES IN EACH PREDICTED VS. FIELD-ASSESSED HAZARD CATEGORY.

Whenever possible, results should be translated into terms that are actionable within existing GMP frameworks. Mapping

predictions onto the hazard categories and response levels already used within the GMP bridges this gap. When evaluating the watercourse PoF model, in addition to reporting a standard predicted-actual scatterplot of PoF values, test results are presented in terms of order-of-magnitude hazard categories as shown in Figure 5. This frames model performance in the decision vocabulary that operators running GMPs in Cambio already use and enables direct comparison with field-inspected hazard classifications.

5.3 Operational deployment of ML/AI

Across all three examples, risk reduction occurs only when model outputs are embedded in workflows with defined decision gates and response pathways. Screening outputs must connect to explicit actions such as initiating detailed assessment, escalating monitoring frequency, deploying field reconnaissance or instrumentation, implementing operational controls, or mitigating. Programs that deploy ML as ‘analytics’ without a corresponding action framework risk creating additional data streams without improving outcomes.

The most valuable deployments of ML/AI operationalize the outputs in support of accelerating GMP program maturity. For example, the LCD model could be used to accelerate programs from reactive/ad hoc (Maturity Level 0) to inventory-based (Maturity Level 1). Watercourse crossing PoF can quickly take an inventory-based hydrotechnical program to systemic/risk-informed (Maturity Level 2). The IMU ML model can accelerate a program rapidly to Maturity Level 3 or even 4 if applied as a complete workflow.

Adoption barriers for ML/AI in pipeline integrity are often organizational rather than technical. Common concerns include model transparency, limited or imbalanced training data, and the perception that ML cannot be regulator-defensible. In practice, these risks can be managed by restricting ML to well-defined screening tasks, maintaining human-in-the-loop review, documenting validation and domain of application, and integrating model outputs into established engineering workflows. When deployed in this manner, ML functions primarily as a triage layer that improves screening consistency and reduces interpretation workload while preserving engineering accountability. For generative AI, the most appropriate applications are knowledge management tasks such as summarizing inspection records, extracting structured data from reports, and supporting documentation consistency, with all outputs reviewed by engineers.

A practical consideration underlying all three case studies is the level of effort required to build production-grade ML models. Even with the assistance of modern AI coding tools, model development remains costly, data-intensive, and methodologically demanding, requiring a diverse team of data scientists, ML engineers, and SMEs. The level of effort and volume of high-quality, labelled training data required often exceed what any single operator can assemble on their own, making scalable solutions that can be deployed across many operators the most practical path to realizing these benefits.

6. ROADMAP FOR OPTIMIZED GMPs

The transition from an interpretation-limited GMP to an optimized, proactive system requires moving beyond isolated data processing toward continuous data integration (Figure 1). In this future state, individual data streams, including lidar, InSAR, and IMU, as well as data and pipeline features identified from other ILI tools, optical satellite imagery, site instrumentation, and historical inspection reports, are continually ingested and assessed by targeted ML models as soon as new data becomes available.

Rather than reviewing these datasets in silos, the results should be synthesized to create a holistic assessment of what, if anything, has changed at a specific site. This integrated approach serves two critical functions: (1) Intelligent Alerting: It draws immediate attention to important changes by cross-referencing anomalies across multiple datasets, thereby increasing confidence in the alert and filtering out single-sensor noise. (2) Streamlined Expert Assessment: It provides the necessary data synthesis to accelerate human review. Expert engineers can easily determine what combination of inputs triggered an alert and intuitively visualize and analyze the underlying source data to arrive at their own engineering determination.

Crucially, this optimized GMP must function as a continuous learning loop. The results of the expert assessments must be captured and fed back into the system to iteratively improve the underlying ML models. For example, if a model identifies landslide-related movement within new LCD results and issues an alert, but the reviewing engineer determines the signature is a false positive (e.g., related to agricultural activity), that determination must be captured to refine the landslide activity model.

Furthermore, any resulting operational actions must be systematically recorded. This not only ensures that required mitigations are executed, but it also aids in subsequent data interpretation. For instance, if landslide activity is detected and a risk control is subsequently employed, the system should expect and contextualize the effects of that mitigation (e.g., earthworks or grade changes) when analyzing the next cycle of lidar acquisitions. Ultimately, this creates a resilient, ever-improving GMP where ML/AI amplifies, rather than replaces, human engineering expertise.

7. CONCLUSION

The examples presented in this paper demonstrate that the primary value of ML/AI in pipeline geohazard management lies in enabling scalable interpretation of monitoring datasets that increasingly define modern GMPs while preserving human engineering oversight. The three case studies demonstrate how ML models can (i) reduce false positives and accelerate triage of IMU bending strain features, (ii) generate consistent inventories of active landslide movement from LCD, and (iii) provide probabilistic screening estimates of watercourse crossing PoF that focus field inspections where they add the most value. Across applications, successful deployment requires validation, high-quality training data, explicit treatment of class imbalance and predictive uncertainty, clear domain-of-use, and integration

into existing risk management workflows with human accountability for decisions. These considerations are particularly important in geohazard applications where complex natural systems and large uncertainties are inherent to the problem. Viewed through the maturity model presented in this paper, ML/AI functions primarily as an accelerator of program development by removing interpretation bottlenecks associated with high-volume monitoring datasets. By automating early-stage screening and interpretation tasks, these tools allow integrity teams to focus engineering effort on the subset of signals most likely to represent credible threats.

In this sense, ML/AI should be understood not as a replacement for established geohazard engineering practices, but as enabling technology that allows integrity teams to scale those practices across increasingly data-rich monitoring environments. As pipeline GMPs continue to mature, responsible scaling of ML/AI offers a pathway to earlier interventions, improved consistency across large pipeline systems, and more efficient allocation of limited engineering and mitigation resources.

ACKNOWLEDGEMENTS

The authors of this paper wish to formally acknowledge the valuable contributions of the Cambio community of pipeline operators, BGC engineers and scientists, as well as the Cambio Earth software developers, data scientists, and product managers. Specific thanks to Michael Porter, Owen Bunce, Patrick Grover, Vincenzo Coia, Joel Van Hove, Caio Stringari, Alex Mitchell, Devin Frioud, Rodrigo Portela, and Katherine Liu for their work on the ML models described in this paper.

Generative AI tools assisted with the preparation of this paper. The following tools were used to provide feedback on draft work, edit text, and improve readability: OpenAI's GPT-5.3 and Anthropic's Claude Desktop (v1.1.5368 with Opus 4.6 and Sonnet 4.6). Cursor (v2.6.14 with Anthropic's Opus 4.6 and Sonnet 4.6, OpenAI's GPT-5.2, and Cursor's Composer 1.5) was used to support coding, data processing, and machine learning. All outputs were thoroughly edited and reviewed. The authors take responsibility for all content.

REFERENCES

- [1] Newton, S., Zahradka, A., Ferris, G., and Porter, M., 2020, "USE OF A GEOHAZARD MANAGEMENT PROGRAM TO REDUCE PIPELINE FAILURE RATES," *Pipeline Integrity Management Under Geohazard Conditions (PIMG)*, ASME Press, pp. 133–137. https://doi.org/10.1115/1.861998_ch14.
- [2] Newton, S., Van Hove, J., Porter, M., and Ferris, G., 2022, "Assessing Geohazard Probability of Pipeline Failure: Lessons and Improvements From the Last 10 Years," *Volume 1: Pipeline Safety Management Systems; Project Management, Design, Construction, and Environmental Issues; Strain-Based Design and Assessment; Risk and Reliability; Emerging Fuels and Greenhouse Gas Emissions*, American Society of Mechanical Engineers, Calgary, Alberta, Canada, p. V001T07A031. <https://doi.org/10.1115/IPC2022-87319>.

- [3] Porter, M., Quinn, P., and Barlow, P., “Conceptual Landslide Velocity Transition Models for a Range of Landslide Behaviour Types.”
- [4] Van Hove, J., Dowling, C., and Barlow, P., 2024, “The Use of Inertial Measurement Unit Data for Managing Slow-Moving Landslides,” *36th International Pipeline Pigging and Integrity Management Conference (PPIM 2024)*, Clarion Technical Conferences, Houston, TX, USA, pp. 539–560. <https://doi.org/10.52202/072781-0030>.
- [5] McKenzie-Johnson, A., and Dotson, R., 2026, “A Deep-Dive into the PHMSA Significant Incident Database and What It Tells Us About Trends in Pipeline Geohazards.”
- [6] Van Hove, J., and Newton, S., 2025, “Operational Experience with Monitoring Tools for Managing the Threat to Pipelines from Slow-Moving Landslides.”
- [7] Maslej, N., 2025, “Artificial Intelligence Index Report 2025,” Artificial Intelligence.
- [8] Zahradka, A., Coia, V., Grover, P., Mitchell, A., and Adair, T., 2025, “Applying Machine Learning to Geohazard Management: Efficient Assessment of Pipeline Watercourse Geohazards by Learning from 3 Million Kilometre-Years of Geohazard Field Inspections.”
- [9] Zahradka, A., Newton, S., Bunce, O., Hove, J. V., and Kowis, J., 2026, “Automated Landslide Impact Detection Through Machine Learning Analysis of Pipeline IMU Bending-Strain Data.”
- [10] Theriault, B., Hart, J. D., McKenzie-Johnson, A., and Paulsen, S., 2020, “Correlation of Single-Run ILI IMU Bending Strain Features to Geohazard Locations,” *Pipeline Integrity Management Under Geohazard Conditions (PIMG)*, ASME Press, pp. 293–301. https://doi.org/10.1115/1.861998_ch31.
- [11] McKenzie-Johnson, A., Theriault, B., Dotson, R., Hart, J., and Varela, P., 2024, “Correlation of IMU Bending Strain Features to Geohazard Locations: An Update,” *Volume 4: Geohazards Management and Strain-Based Design and Assessment*, American Society of Mechanical Engineers, Calgary, Alberta, Canada, p. V004T06A008. <https://doi.org/10.1115/IPC2024-133093>.
- [12] Dowling, C., Barlow, P., Van Hove, J., Hanvi, T., and Hart, J., 2024, “Ground Movement Or Construction? How to Identify Ground Movement Signatures in Inertial Measurement Unit Data,” *36th International Pipeline Pigging and Integrity Management Conference (PPIM 2024)*, Clarion Technical Conferences, Houston, TX, USA, pp. 1339–1360. <https://doi.org/10.52202/072781-0075>.
- [13] Scheevel, C., Dowling, C., Hart, J. D., and Cook, D., “IMU Bending Strain: Analysis as Geohazard Screening Tool.”
- [14] Dewar, D., 2020, “Incorporating Inline Inspection Internal Measurement Unit Data Analysis Into Integrity Management Programs.”
- [15] Zahradka, A., Newton, S., Scheip, C., Bunce, O., Stringari, C., and Kowis, J., 2026, “USE OF MACHINE LEARNING TO IDENTIFY LANDSLIDE SIGNATURES IN IMU BENDING STRAIN DATA [UNDER REVIEW],” *Proceedings of the 2026 16th International Pipeline Conference*.
- [16] Lato, M., Van Veen, M., Ferrier, A., Weidner, L., and Graham, A., 2022, “Predicting the Future by Mapping the Past: Revolutionary Innovations in Lidar Change Detection Analysis Are Enabling Regional Scale Mapping and Identification of Threats From Geohazards,” *Volume 3: Operations, Monitoring, and Maintenance; Materials and Joining*, American Society of Mechanical Engineers, Calgary, Alberta, Canada, p. V003T04A012. <https://doi.org/10.1115/IPC2022-87104>.
- [17] Weidner, L., Ferrier, A., Van Veen, M., and Lato, M. J., 2024, “Rapid 3D Lidar Change Detection for Geohazard Identification Using GPU-Based Alignment and M3C2 Algorithms,” *Can. Geotech. J.*, **61**(5), pp. 896–914. <https://doi.org/10.1139/cgj-2023-0073>.
- [18] Anderson, R., Lukas, V., and Aichele, S. S., 2024, *The 3D National Topography Model Call for Action—Part 1. The 3D Hydrography Program*, 1519, Reston, VA. <https://doi.org/10.3133/cir1519>.
- [19] Meinshausen, N., “Quantile Regression Forests.”
- [20] Yin, X., Fallah-Shorshani, M., McConnell, R., Fruin, S., Chiang, Y.-Y., and Franklin, M., 2023, “Quantile Extreme Gradient Boosting for Uncertainty Quantification.” [Online]. Available: <http://arxiv.org/abs/2304.11732>. [Accessed: 30-Oct-2024].
- [21] Chen, T., and Guestrin, C., 2016, “XGBoost: A Scalable Tree Boosting System,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, San Francisco, California, USA, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
- [22] Bozorgzadeh, N., and Feng, Y., 2024, “Evaluation Structures for Machine Learning Models in Geotechnical Engineering,” *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, pp. 1–8. <https://doi.org/10.1080/17499518.2024.2313485>.