Using Machine Learning to Manage SCC: Data Strategy and Best Practices

Syed Aijaz[1], Clifford Maier[1], Michael Gloven[2]

[1]TC Energy, [2]Pipeline-Risk (PLR)

Organized by

CLARION
TECHNICAL CONFERENCES

https://doi.org/10.52202/078572-0012

https://doi.org/10.52202/078572-0012

Abstract

Big data, machine learning and artificial intelligence: these buzzwords invariably create a lot of hype but do they actually live up to it? TC Energy's (TCE's) threat management team set out to answer that question by curating a comprehensive dataset of pipeline integrity digs in-ditch findings to create a machine learning model for predicting stress corrosion cracking (SCC) on its transmission pipeline assets. In collaboration with Pipeline-Risk (PLR), in-ditch results from 1800+ digs since 2012 were consolidated. A classification-based machine learning model was trained on a subset of the consolidated dig data, and its performance was then evaluated using the remaining, unseen portion of the dataset. The results exhibited a predictive efficacy of 87+% for predicting likelihood of SCC on TCE transmission pipelines. This model is now utilized in assisting in SCC direct assessment (SCCDA) dig site selection workflows and prioritizing in-line inspection (ILI) assessments resulting in significant efficiency gains. The most important factors contributing to the success of this model were an organized enterprise data infrastructure, subject matter experts (SME)-driven criteria and integrated team structure allowing a comprehensive threat management approach. This paper showcases the results from this project, and provides recommendations to ensure your organization is optimally positioned with people and processes to maximize value created by the computational prowess of machine learning.

How have we historically been performing SCC threat management?

Effective management of any pipeline integrity threat begins with a good understanding of that threat. The pipeline industry already had some understanding of SCC by the early 1990s, when Canada's National Energy Board (NEB) launched inquiries into this cracking phenomenon that had produced several failures. Recommendations from the NEB inquiry report[1] eventually led to regulations that required Canadian pipeline operators to develop and implement an SCC management program.

Pipeline operators outside of Canada generally had SCC management programs in place by the time the NEB inquiry report was published. It is clear, however, that the report promoted further development of these programs. Additionally, SCC research activity intensified across the industry following the NEB inquiry report. This activity resulted in an improved understanding of the cracking mechanism and additional guidance for pipeline operators.

ASME B31.8S Managing System Integrity of Gas Pipelines was updated in 2004 to provide guidance for SCC management. Additional guidance was provided by other industry publications such as NACE RP0204 Recommended Practice Stress Corrosion Cracking (SCC) Direct Assessment Methodology (2004), ASME STP-PT-011 Integrity Management of Stress Corrosion Cracking in Gas Pipeline High Consequence Areas (2008), and CEPA Recommended Practice for Managing Near-Neutral pH Stress Corrosion Cracking, 1st Edition (2011). Some of these publications have since been updated.

https://doi.org/10.52202/078572-0012

Many operators have developed SCC management programs that utilize susceptibility criteria based on these industry documents. In applying the criteria to specific pipeline segments, operators often consider broad characterizations about operating stress level, coating type, and age of the pipeline that may not accurately reflect localized conditions. Additionally, soil and terrain characteristics and climate conditions are sometimes ignored or poorly integrated. SCC is a complex phenomenon with various interdependent causal factors. Consequently, generalized and broad treatment of data may result in non-specific results that do not adequately account for local variability in causal parameters. Ineffective SCC Direct Assessment (SCCDA) programs, having a limited understanding of crack detection in-line inspection results, and inappropriate prioritization of pipeline segments are evidence of such an approach. Outcomes can generally be improved by incorporating more detailed and localized information that is specific to the pipeline segment being evaluated. To that end, pipeline operators need to develop a data-driven threat management approach that incorporates field findings, in addition to utilizing susceptibility criteria based on industry recommendations.

How can we move towards a more robust, data-driven approach?

Developing a data-driven threat management methodology does not necessarily entail an increase in the amount of data collected. It's likely that most operators can utilize data that already exist. Namely, operators collect a wealth of information during excavations. Many SCC management programs incorporate a portion of the excavation data collected, with some of the unused data having the potential to unlock additional benefits for SCC management efforts. Thus, operators sometimes miss opportunities to gain insights that could have a material impact on their SCC management program. Excavation data can be supplemented with soil and climate data from publicly available data stores.

Machine learning is a relatively new, data-driven approach for developing integrity threat susceptibility prediction models that can be leveraged by pipeline operators. Data used to develop machine learning models can be pipeline specific and reflect a higher level of detail and localization compared with broad categorizations and generalized assumptions that are more common within SCC management programs. Operators can use machine learning to discover data patterns and interdependencies associated with various susceptibility factors related to SCC findings that would be otherwise difficult to account for using analytical empirical methods. This includes gaining an understanding of how relationships between susceptibility factors influence the likelihood of SCC occurring for a given pipeline system. Such insights have been challenging to recognize historically because legacy SCC management approaches have generally considered the coexistence of susceptibility factors only. With machine learning, complex relationship among many in-ditch observations can be utilized and learned from.

https://doi.org/10.52202/078572-0012

How did TCE utilize in-ditch data to create a machine learning model?

The SCC machine learning model effort was launched in January 2024 and the final model was completed in July 2024. The primary goal was to utilize in-ditch data collected during all pipeline integrity digs and develop a machine learning model to predict the likelihood of finding SCC across the TCE natural gas transmission pipeline system.

The decision to use in-ditch, dig data instead of electro-magnetic acoustic transducer (EMAT) ILI-reported crack features was motivated by the desire to remove the uncertainty associated with the probability of identification (POI) and probability of detection (POD) characteristic of EMAT sensors. ILI-reported features and field-reported features may differ and this may cause the training dataset to misinterpret feature call-outs as the target of interest (SCC) rendering the SCC classification error-prone. This can happen if the ILI tool reports a crack and upon in-ditch investigation it turns out to be a different type of feature. This outcome would introduce errors into the training dataset. Conversely, in-ditch observations during a dig (using non-destructive examination (NDE) methods) are a confirmation of the feature and feature type and are therefore definitive. For this reason, the TCE SCC team curated a dataset comprising of exclusively in-ditch observations. By consolidating several categories of observations (shown in table 2), the team attempted to replicate the dig site conditions to predict likelihood of SCC.

Creating the dataset best describing SCC conditions

SCC is a form of environmentally assisted cracking that occurs when a combination of environmental conditions and tensile stresses act coincidently on a pipeline made from a susceptible steel, that results in the formation of cracks on the surface of the pipe. Hence, the susceptibility factors can be boiled down to three essential components: the pipeline material, environmental conditions surrounding the pipe joint, and the active or residual stress on the pipe.

To best align the training dataset to these susceptibility factors the dataset was curated to reflect these data categories. The initial sample data was comprised of 1859 individual digs from 2012-2023. These digs were driven by a variety of factors such as post-ILI digs or direct assessment digs. The digs for which the data did not meet the standard of completeness and quality for machine learning modelling were not used, resulting in a final sample size of 1827 individual digs. Each individual dig site exposed up to a single joint of pipe. Table 1 shows the dig data drivers.

https://doi.org/10.52202/078572-0012

Table 1. Dig project drivers

Total digs	1859
Digs used for machine learning	**1827**
Post MFL ILI digs	1017
Post EMAT ILI digs	388
SCCDA digs	188
ECDA digs	48
Material verification digs	31
Other digs	155

These digs represent a significant proportion of the geographic footprint of TCE's US gas transmission pipeline assets and therefore provide a representative sample. This includes piggable pipelines where TCE had completed an ILI inspection, as well as non-piggable assets where other assessment methods were utilized such as direct assessment.

Several data sources containing in-ditch observations collected during various integrity digs were identified to create the training dataset, seen in table 2.

Table 2. Machine learning training dataset – TCE data

Data Category	Field
Defect Information	Defect type
	Crack type
	Crack length
	Crack depth
	Crack max percent depth
	Coincident metal loss
Pipeline Coating	Coating type
	Brand
	Condition as found
	Application method
	Application year
	Defect type
Pipeline Properties	Pipe OD
	Nominal wall thickness
	Pipe grade
	MAOP
	Operating percent SMYS
	Install date
	Manufacturer
	Longitudinal seam type

https://doi.org/10.52202/078572-0012

Soil and Terrain	Land usage
	Soil pH
	Soil type
	Soil texture
	Fragment size
	Depth of cover
	Drainage quality
	Soil resistivity
	Site topography
Geospatial	Dig site GPS
	Pipeline name/system
	Distance to nearest compressor station

In addition to TCE data sources, PLR included publicly available data sources to further augment the training dataset including SSURGO, PRISM and NRI data. These are publicly available datasets that provide valuable soil, terrain, and climate information for US.

SSURGO[2] - The SSURGO database contains information about soil (such as soil texture, properties, drainage, electrical conductivity, soil reactions etc, frequency of flooding etc.) as collected by the National Cooperative Soil Survey over the course of a century. The information can be displayed in tables or as maps and is available for most areas in the United States and the Territories, Commonwealths, and Island Nations served by the USDA-NRCS. The information was gathered by walking over the land and observing the soil. Many soil samples were analysed in laboratories. And collected at scales ranging from 1:12,000 to 1:63,360.

PRISM[3] - The PRISM Climate Group gathers climate observations from a wide range of monitoring networks, applies sophisticated quality control measures, and develops spatial climate datasets to reveal short- and long-term climate patterns.

National Risk Index (NRI)[4] - The National Risk Index is a dataset and online tool to help illustrate the United States communities most at risk for 18 natural hazards. It was designed and built by FEMA in close collaboration with various stakeholders and partners in academia; local, state and federal government; and private industry.

Refining and enhancing the dataset for machine learning

A few key enhancements were performed on the training dataset to maximize the value from the dig findings.

https://doi.org/10.52202/078572-0012

Distance to nearest compressor station

For digs on a pipeline with a compressor station located upstream or downstream, the conservative assumption was made to use the distance to the nearest compressor station regardless of flow direction. This was to account for any bi-directional lines where the flow direction changes or may have changed in the past. This field was added to be in alignment with ASME B31.8S data collection recommendations for SCC threat assessment.

Pipe coating brand

Coating brand was added as a separate field to account for Dearborn wax coating. This type of coating was a key contributing factor on one TCE in-service SCC-related failure. This coating is typically found to be highly degraded upon excavation and is commonly found to be associated with and SCC colonies. However, the coating brand data was sparsely available. Therefore, coating type field was modified to account for Dearborn wax as a coating type distinct from other types of wax coating.

Soil texture

On dig sites where the soil was found to contain any amounts of clay or clay-like texture regardless of other soil types found, the soil type was assumed to be clay. This was to account for the detrimental effects of clay's expansion and contraction due to varying levels of moisture on the pipe coating.

Supervised Machine Learning Process

The process of developing a machine learning model relies on the iterative refinement of an initial model by training the model on observed outcomes. One or more available machine learning models can be deployed, tested and validated on subsets of this training data and the best performing model is selected based on its performance. A model's performance is measured by the metrics of accuracy sensitivity, specificity, and balanced accuracy.

Accuracy

This is defined as the percentage of true calls (true positives and true negatives)

$$Accuracy = \frac{(True\ positives + True\ negatives)}{(True\ positives + False\ positives + True\ negatives + False\ negatives)}$$

Sensitivity

This is defined as the percentage of correctly predicted calls of actual positive instances

$$Sensitivity = \frac{(True\ positives)}{(True\ positives + False\ negatives)}$$

https://doi.org/10.52202/078572-0012

Specificity

This is defined as the percentage of correctly predicted calls of actual negative instances

$$Specificity = \frac{(True\ negatives)}{(True\ negatives + False\ positives)}$$

Balanced Accuracy

This is defined as the average of sensitivity and specificity

$$Balanced\ Accuracy = \frac{(Sensitivity + Specificity)}{2}$$

The project followed an iterative supervised machine learning process as shown in figure 1. For each dig, a binary value was defined as the outcome based on whether SCC was observed or not. The input variables were defined based on SME input and availability of quality and complete data. The objective of the process was to use machine learning methods to utilize underlying patterns within the predictor data obtained from visual inspections. These patterns became the basis of candidate SCC prediction models.

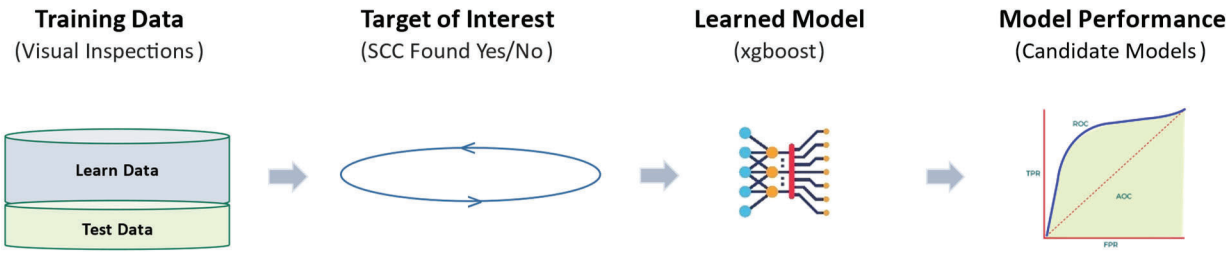

Figure 1. Machine learning process

Training data was divided into learn and test data. The learn data was used to learn candidate models based on iterating through different machine learning algorithms and their tuning parameters. Test data was used to validate the performance of candidate models.

As seen in table 3, there were various underlying drivers for the digs performed by TCE. The learning and test data was divided on basis of the type of dig. All digs that were driven by an EMAT feature call that required investigation were called the “SCC digs”. The based on all other factors (post MFL ILI dig, ECDA, SCCDA, material verification) were classified as “non-SCC” digs. Henceforth, the machine learning algorithm was trained on the non-SCC digs and its performance tested on the SCC digs.

https://doi.org/10.52202/078572-0012

Table 3. Dig dataset type

Digs used for machine learning	1827	Dig type	Dataset Type
Post MFL ILI digs	1017	Non-SCC	Learn data
Post EMAT ILI digs	388	SCC dig	Test data
SCCDA digs	188	Non-SCC	Learn data
ECDA digs	48	Non-SCC	Learn data
Material verification digs	31	Non-SCC	Learn data
Other digs	155	Non-SCC	Learn data

This was done for one main reason: all dig sites labelled "SCC digs" were driven by existing knowledge that there might be SCC found at that location, due to the EMAT-reported feature. As a result, these SCC dig records would introduce an inherent bias and lead to an overperforming model. To prevent this, the model was trained on all dig sites where integrity teams did not have prior knowledge of SCC being found, or where the objective was to remediate non-SCC features (metal loss, manufacturing etc.) and in the process found SCC. In early iterations of the project, this was found to be true and thereafter, the learn/test data split was modified.

The learning target of interest was based on SCC being found or not found during the in-ditch inspection, hence the project followed a binary True/False classification learning process where the resulting candidate classification models output a probability of SCC. Models were applied to the test data to assess performance and learning curves supported selection of a final model as shown in figure 2.

Candidate models were learned based on selected machine learning methods including XGBoost, random forest, linear regression and logistic regression. Models were tested for accuracy (% of correct calls), specificity (% of correct 'No SCC' calls) and sensitivity (% of correct "Yes SCC' calls). In the final analysis, XGBoost performed best as far as sensitivity and the method's ability to manage missing data, predictor non-linearities and interactions.

https://doi.org/10.52202/078572-0012

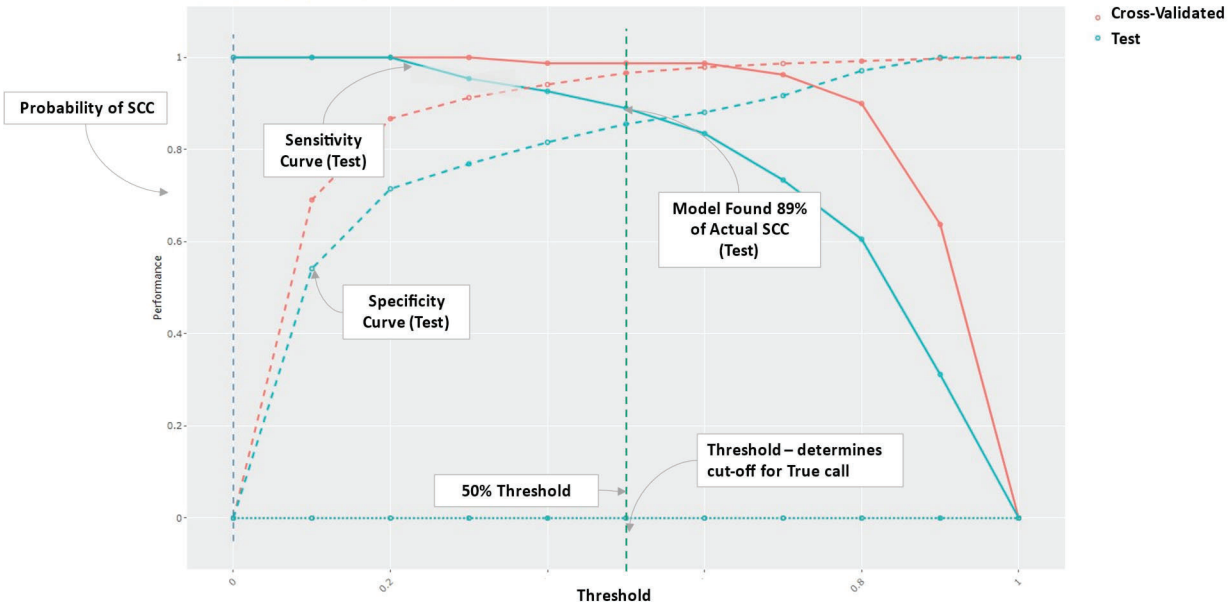

Figure 2. Model performance learning curves

Learning curves are useful in determining whether a resulting model meets acceptance criteria for use in making predictions by visualizing the performance metrics. The model's sensitivity and specificity performance curves are plotted against a threshold for predicting true or false. Adjusting the threshold values can impact the model's performance:

- Sensitivity Curve: Typically, as the threshold decreases, sensitivity increases because the model becomes more lenient in classifying SCC. However, this may come at the cost of lower specificity.
- Specificity Curve: Conversely, as the threshold increases, specificity improves because the model becomes stricter, reducing false positives but potentially missing true SCC cases.

https://doi.org/10.52202/078572-0012

Model insights were gained through global and local predictor explainability analysis. Figure 3 shows predictor importance values of candidate models providing a global view of the overall ranking and weight of individual predictors.

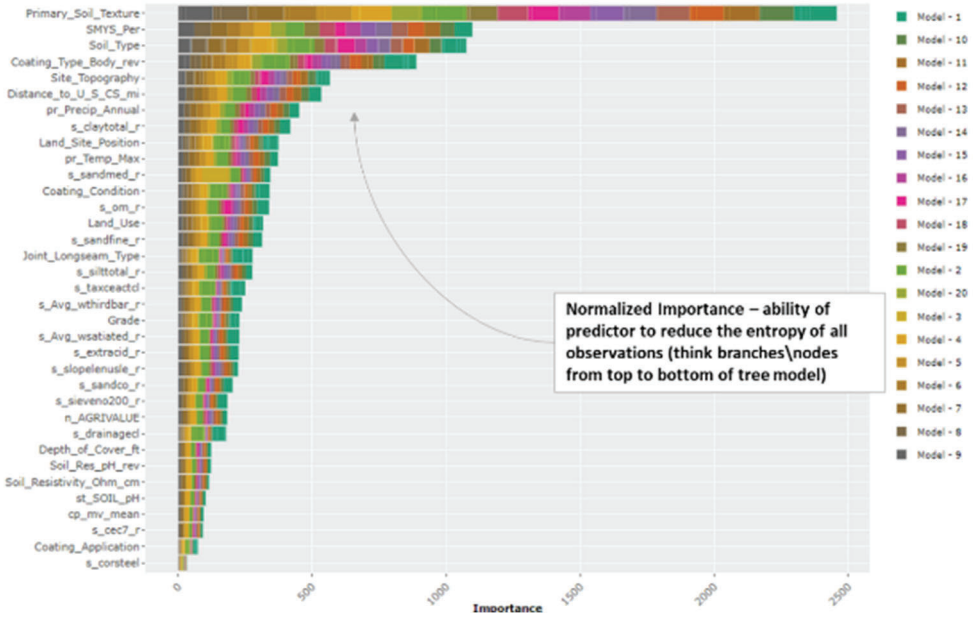

Figure 3. Predictors importance values

Figure 4 shows deconstructed predictions on joint or dig basis illustrating the non-linear and interactive nature of individual predictors on the local prediction. Each bar and its stacked segments represents the predictors contribution to the prediction of SCC being found or not.

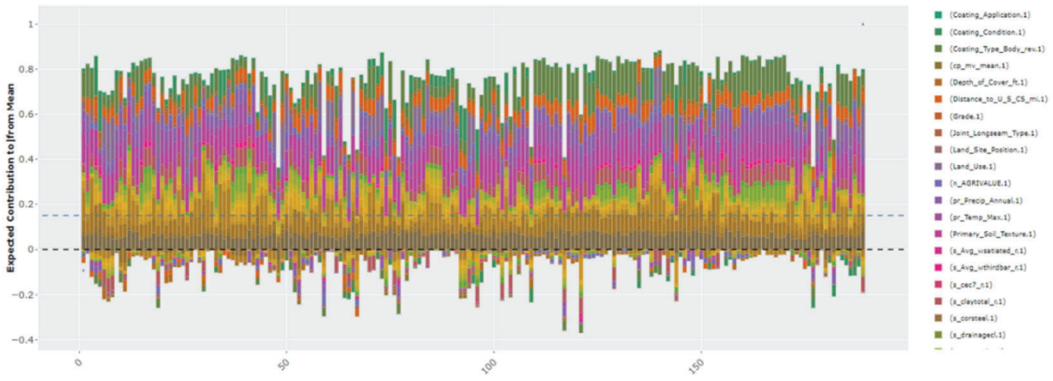

Figure 4. Prediction results per pipe joint

https://doi.org/10.52202/078572-0012

Results

Using these performance evaluation metrics, the TCE machine learning model results can be seen in table 4 and table 5. The model performance resulted in a balanced accuracy of 87% on the test data. This percentage signifies the model's prediction correctness for whether SCC is found or not found on unseen, test data.

Table 4. Model prediction results

	Predicted Negative	Predicted Positive
Actual Negative	239	40
Actual Positive	12	97

Table 5. Model performance metrics

Evaluation Metric	Value	
Accuracy	87%	
Sensitivity	89%	
Specificity	86%	
Balanced Accuracy	87%	

The top ten predictors of SCC according to the model are seen in table 6. These predictors are mostly consistent with industry guidance and documents, as reflected in operating percent SMYS, distance to compressor station, and coating type. Soil and climate fields also appear to be strong predictors of finding SCC, and can be explained by their detrimental effects on the coating condition. For instance, clay soil in an undulating to depression topographical area with low annual precipitation may result in clay compaction over time, contributing to forces exerted on the coating system and potentially damaging it.

Table 6. Model top 10 predictors of finding SCC

Top 10 machine learning predictors for SCC
Soil texture
Operating percent SMYS
Soil type
Coating type
Distance to nearest CS
Site topography
Annual precipitation
Clay total
Maximum temperature
Silt total

https://doi.org/10.52202/078572-0012

What does this all mean?

These results provide validation of the proof of concept that machine learning can be utilized to derive valuable and reliable predictions for integrity threats. TCE's SCC team continues to improve and develop the machine learning model by enhancing predictors, addressing completeness and quality concerns and incorporating additional dig datasets. Attaining an 87+% predictive efficacy as a first pass shows promise for further improvements to model performance.

Therefore, machine learning as an addition to pipeline integrity threat management processes can demonstrably enhance an operator's threat management program. Presently, the SCC team is utilizing these results to support in the annual SCCDA dig site selection process to improve effectiveness of the dig program, and to prioritize EMAT inspection segments on a transmission pipeline system in the US.

It should be noted that like any statistical prediction model, an underlying principle is that the data used for developing the model is representative of the population on which the predictions are made. Consequently, it is conceivable that conditions may exist which differ significantly from the training dataset and therefore, the expected accuracy of the model predictions may suffer. The authors will continue to examine the efficacy and performance of the model to further establish the performance of the model over a wide range of conditions.

What factors contributed to the success of this machine learning effort?

The most important contributing factor to the model's performance is the quality data available for a large sample of dig locations. This was enabled by TCE's implementation of key data infrastructure enhancements that have enabled leveraging value for the entire organization. The TCE data lake is a repository of all key enterprise data sources that were previously disparate and disconnected and now are in one convenient location. These data sources include, but are not limited to, GIS data, ILI results, CP surveys, soil and terrain surveys, and pipe properties. The implementation of this cloud-based data infrastructure has enabled access to a vast array of data and allowed multiple teams' data to be connected to reveal valuable synergies. This was instrumental in constructing a quality dataset to train the machine learning model.

TCE leaders have cultivated a culture that encourages the development and implementation of new ideas. This emphasis on innovation has helped retain skilled data professionals who have made significant advancements in data engineering, management, and tool development for TCE data. These efforts have improved the quality, completeness, and integration of key data sources.

Additionally, the structured process for handling data collected during digs plays a crucial role. The entire workflow for collecting data from field reports, on-site imagery, visual inspections, and NDE

https://doi.org/10.52202/078572-0012

results is managed by specialized teams. These teams, which include subject matter experts, integrity engineers, and data engineers, thoroughly review and validate the data. Once vetted, the dig results are finalized and published.

Implementing machine learning into an integrity management program requires two key components: a quality data infrastructure, and a team of skilled data practitioners dedicated to developing data-driven tools and solutions. To this end, operators can utilize existing data sources, historical records, and reports to complete and refine that data, and integrate it with publicly available soil and climate data sources. Further, training and enriching teams with data experts can elevate the quality of data available for machine learning. Implementing these changes will support the pursuit of machine learning opportunities to enhance threat management programs. These statistically driven models can complement an operators existing risk program, SCCDA methodologies and provide general guidance for prioritizing future integrity assessments.

https://doi.org/10.52202/078572-0012

References:

1. Canadian Government Publishing Centre. "Pipeline Stress Corrosion Cracking." Available at: https://publications.gc.ca/collections/Collection/NE23-58-1996E.pdf.
2. Natural Resources Conservation Service. "Soil Survey Geographic Database (SSURGO)." United States Department of Agriculture. Available at: https://www.nrcs.usda.gov/resources/data-and-reports/soil-survey-geographic-database-ssurgo
3. PRISM Climate Group. "PRISM Climate Data." Oregon State University. Available at: https://prism.oregonstate.edu/
4. Federal Emergency Management Agency. "National Risk Index: Learn More." Available at: https://hazards.fema.gov/nri/learn-more

https://doi.org/10.52202/078572-0012