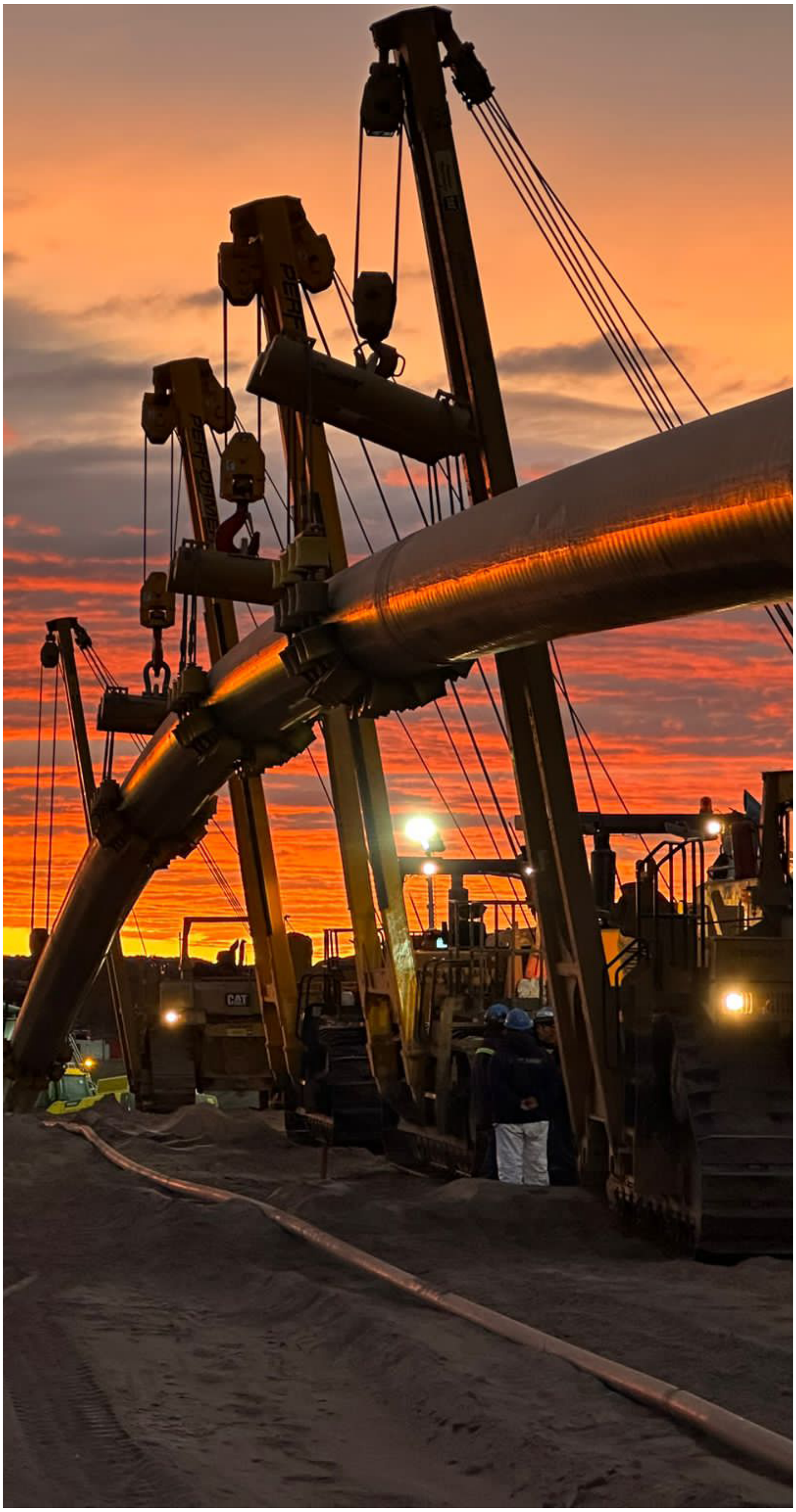
4/2024

Pipeline Technology Journal



Construction & Maintenance

WWW.PIPELINE-JOURNAL.NET

e-ISSN 2196-4300 / p-ISSN 2751-1189

Using Data Analytics to Unearth the Relationship Between Geohazards and Pipeline Corrosion

J. WALKER, E. BENNETT, J. MARTIN, J. SOLTIS > ROSEN GROUP

Abstract

Prediction of external corrosion using data from in-line inspections (ILI) can be instrumental in making proactive pipeline integrity management decisions, especially when integrated with geospatial data. This work investigates how a relationship may be established between the presence of bending-strain areas and the initiation and or progression of external corrosion. As the initiation of external corrosion requires both failure of the external coating and failure of the cathodic protection system, it may be possible to limit the damage caused in areas of bending strain by adopting inspection and repair strategies reflecting the pipeline integrity risks. The study examines how integrating geohazard characteristics into machine learning models improves the prediction of external corrosion. The paper provides insights into the complex interplay of geohazards, bending strain, and corrosion, informing better integrity management strategies. This study suggests that a shift towards a holistic, automated, machine learning-assisted approach for pipeline integrity can be attained in compliance with regulation and highlights the importance of data-driven approaches for safety and reliability.

1. Introduction

Understanding pipeline corrosion, which initiates from failures in protective mechanisms like coating defects or insufficient cathodic protection, is critical for the operation and integrity management of oil and gas pipelines. Recent studies have shown the effectiveness of machine learning algorithms in this area [1] [2] [3]. External environmental factors introducing bending strain to a pipeline can lead to significant deviations in its expected operating conditions, resulting in an increased and potentially unacceptable level of risk. While the applied stresses can cause tensile fracture at girth welds and local buckles due to compressive loads, the external protective coating can also be damaged even if the overall integrity of the pipeline is considered acceptable for continued operation. A damaged pipeline coating inherently leads to an increased risk of corrosion. Natural causes that can lead to external loads on a pipeline include precipitation, earthquakes, karst subsidence, and frost heaving, to name a few. Although bending strain is primarily considered a pipeline integrity-related issue, inertial measuring units (IMU) are included in a greater and greater proportion of in-line inspection tools. If a relationship can be appropriately modeled between the risk of external corrosion and the presence of bending strain, there is the possibility of an additional value-adding from including an IMU (beyond location tracking) for pipelines typically considered low-risk in relation to bending strain failure.

There are many mechanisms and morphologies of external corrosion in buried pipelines, and of those, a subset associated with bending strain includes the following:

1. Pitting Corrosion: Bending strain can cause local thinning of the protective coating and localized breakdown, leading to highly localized pitting corrosion.

2. Crevice Corrosion: Bending and strain can cause slight gaps in coatings or seam welds leading to crevices where corrosive agents may become trapped.

3. Corrosion Fatigue: Cyclic stresses from the environment or product can lead to material fatigue damaging protective coatings and later sustaining this damage mechanism.

4. Stress Corrosion Cracking (SCC): Processes where a combination of tensile stress and a corrosive environment leads to cracking.

In this work, we take a high-level approach to understanding the relationship between external corrosion, the pipeline and its local environment. Rather than constructing models that predict the specific corrosion types or propagation over time, we aim to develop models that provide insights into the most at-risk pipeline sections. This approach allows us to build powerful decision-aid tools focused on determining high-risk pipeline sections and advance our understanding of those geospatial and geo-temporal features that put pipeline infrastructure most at risk. We aim to demonstrate a link between bending strain risk factors and external corrosion. We do this by introducing geospatial and geo-temporal features into our data analysis and predictive models representing the real-world features that pipelines may experience. This selection of environmental features more commonly associated with bending strain may be used to enhance our understanding of external corrosion.

We approach this challenge in two ways.

First, we demonstrate correlations between slope units and pipeline joints with known anomalous features of bending strain and external corrosion over a small-scale study area. Second, we construct a proof-of-concept external corrosion binary classifier and evaluate its performance for different combinations of feature inputs.

2. Integrity Data Warehouse

ROSEN is well-positioned to perform insightful data exploration and develop powerful predictive models relating to pipeline integrity. This is made possible by ROSEN's comprehensive warehouse of pipeline inspection data, the Integrity Data Warehouse (IDW). The IDW contains over 26,000 historical inspections across pipelines located globally and contains detailed pipeline information including routes, product, manufacturer details, and pipeline defect features obtained through in-line inspection (ILI). The IDW has also been enriched with additional geospatial features, which enhance the data provided by ILI by introducing contextual information that can be integrated into predictive models. We can further improve our understanding of defect risk factors by understanding the relationship between geo-enrichment variables and known pipeline sections that contain defects (or sections without defects). We have introduced the following datasets into the IDW:

- Geology: Geological period and mineral composition [4]
- Soil: Type, material content, moisture and pH [5] [6]
- Groundwater: Depth and distance to the nearest water table [7]
- Precipitation: Hourly precipitation data [8]
- Elevation: Digital elevation maps (DEM) constructed from RADAR and LIDAR point cloud data from the 3DEP project [9] [10] [7]
- Terrain Classification: Clustering of terrain into discrete definitions [11] [12]
- Historical Landslide data: Distance and duration since nearby landslides [13] [14] [15]
- Land use classification: Clustering of land use into discrete definitions [16]
- Open Street Maps: Intersections with roads, railways, or waterways [17]

In some cases, these datasets are pre-processed or aggregated in a bespoke manner. For example, the precipitation data is aggregated into percentiles over monthly periods, and LIDAR DEM tiles are processed at a resolution of 1m from point cloud data. These datasets are available to query for any combination of latitude and longitude locations, not just those directly on the pipeline route. This flexibility allows us to collate information on surrounding points of interest within a given proximity to pipelines.

3. Slope Units

We consider a subset of pipelines within the IDW for which we have inspection data sensitive to both external corrosion and bending strain defect features. In

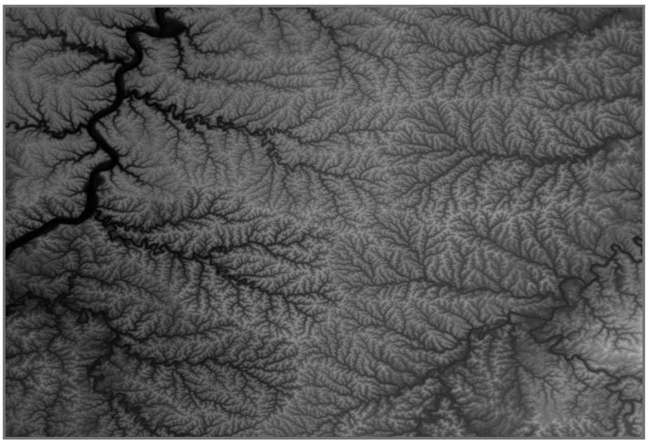


Figure 1: Example RADAR DEM represented by the greyscale colour mapping [7].

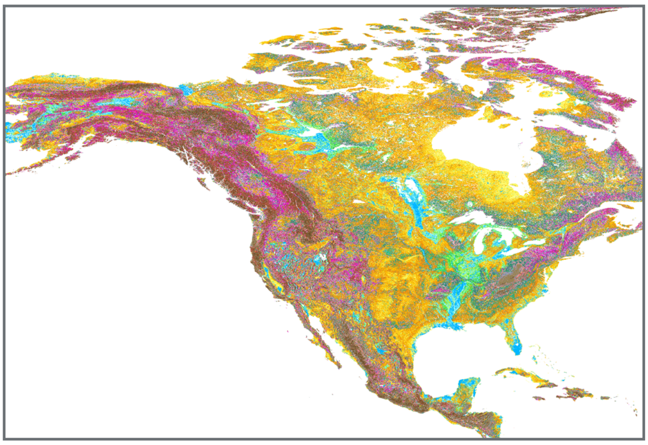


Figure 2: Example Terrain22 classification of continental North America with each color representing a unique terrain type [11] [12].

particular, we consider a region of size 150,000km² within the states of Ohio, West Virgina and Pennsylvania, USA. This region is an insightful study region as it contains pipelines within the Appalachian Mountains and the Interior Lowlands, providing a diverse geography within the dataset. Slope units are segmented regions of terrain partitioned by drainage and divide lines. They provide a natural landscape clustering as they inherently incorporate terrain geometry in their structure, making them powerful tools for geohazard risk assessment. A typical slope unit has a length scale of ~100m, which is a suitable scale for capturing meaningful aggregation of geospatial features. Thus, they are a powerful component in calculating geohazard risks at scale [18].

3.1 SUMak algorithm

Slope units are expensive to compute; however, with careful selection of regions of interest and pragmatic use of resolution scales, their calculation can be performed efficiently. In this work, we have rewritten and deployed a bespoke version of the Slope Unit Maker

(SUMak) algorithm [19] in the Python programming language which has been interfaced with the IDW. Python and the Celery task broker system [20] allow for the computation of these slope units over vast land areas in relatively short order, enabling slope units to be determined at a scale not historically realistic to achieve. Thus, we can interface between geospatial and inspection data and calculate slope units over relevant pipeline regions. The SUMak algorithm is a parameter-free model and, therefore, scale-invariant, which allows us to segment SUMak into two stages:

1. Watershed delineation using RADAR tiles at 30m resolution

2. Slope unit delineation using LIDAR tiles at 1m resolution

This approach allows us to refine our choice of DEM to regions of terrain where points of interest intersect with watersheds before executing the SUMak algorithm at a higher resolution in relevant areas. This two-stage approach thus vastly reduces the overall computation time.

3.2 Geospatial processing

A set of pipelines with examples of bending strain and external corrosion anomalies are selected, and the slope units within a 2 km perpendicular distance along the entire length of the pipeline are calculated using the methodology described above. For each slope unit, we also collect and aggregate several geo-enrichment features outlined in section 3 over each slope unit area. With a dataset of slope units, we can study the relationship between their geospatial features and proximity to pipeline defects.

We construct three region categories of interest:

1. External corrosion regions

2. Bending strain regions

3. Control regions

For external corrosion, a region is defined by the presence of at least one external corrosion feature with a depth greater than 10% wall thickness; for bending strain, the region is defined as a section of pipeline with a maximal strain greater than 0.1% and assessed to be likely caused by external environmental factors. The control regions are adjacent to the aforementioned regions, 500 m upstream/downstream along the pipeline route, where no external corrosion or bending strain features have been reported. Each region is mutually exclusive and the slope units within 500 m of the pipeline route were collected. Figure 3 depicts an example of a bending strain region constructed with this methodology.

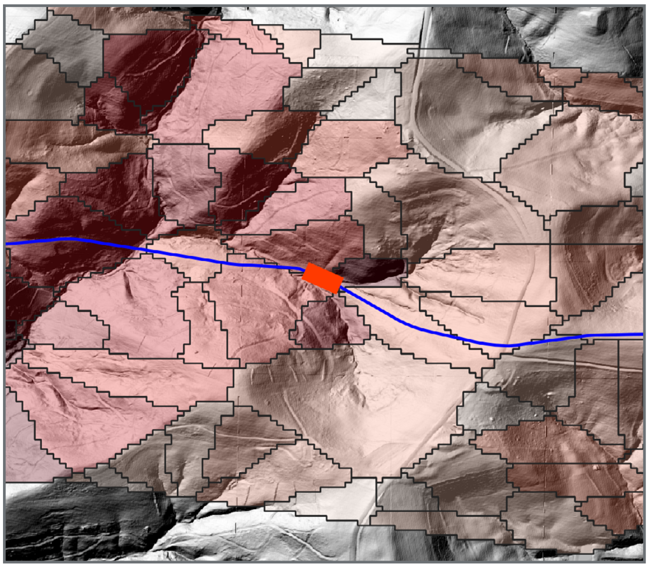


Figure 3: A region of pipeline with bending strain indicated by the solid red line with 500m up and down-stream indicated by the blue line. The surrounding shaded polygons represent slope units with a colormap representing slope (deg). The background shows the hillshade with sun at azimuth 315 degrees and altitude of 45 degrees.

3.3 Results

Based on the above methodology, the pipelines within the study area have been segregated into:

1. 167 external corrosion regions with a total area of 131km²

2. 144 bending strain regions with a total area of 113km²

3. 175 control regions with a total area of 137km²

By taking histograms across these regions, we probe feature importance and underlying correlations which might help us better understand high-risk factors. We also directly compare the distributions of geospatial features for each region; here, we will consider the slope angle. Figure 4 shows that a higher proportion of

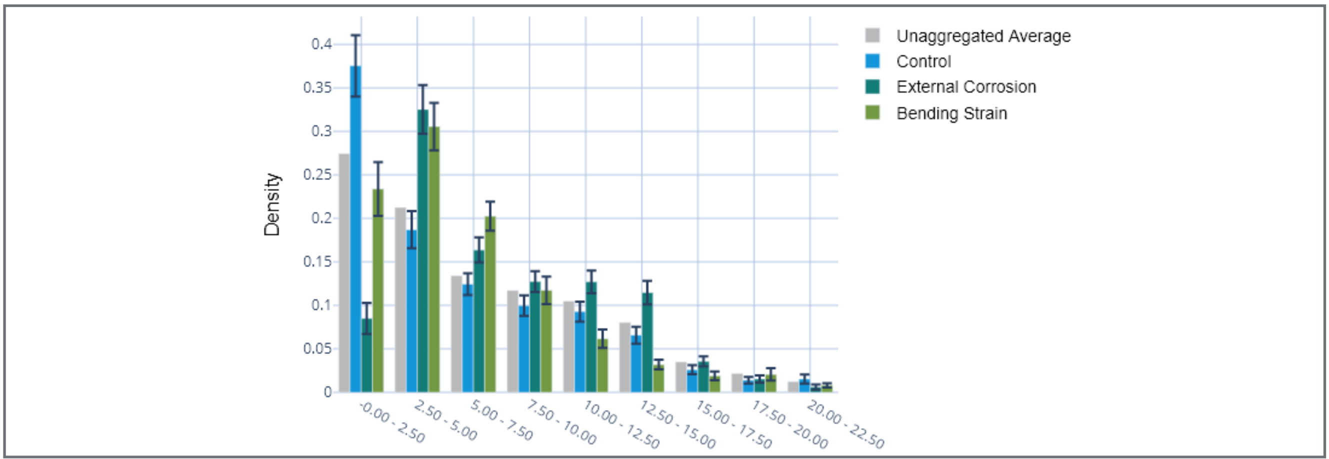


Figure 4: Histogram of the slope angle of slope units over all regions aggregated over region type. An additional histogram Unaggregated Average is the average slope angle of all slope units that exist within the IDW to date. Error bars represent the standard error across each bin over all regions of each category.

external corrosion and bending strain regions contain slope units with higher degrees of slope angle than control regions. We also average over all 20,000 calculated slope units contained within the IDW and see that the control regions have a higher proportion of a slope angle less than 2.5 degrees than this unaggregated average, indicating that the slope angle is a potentially good indicator of risk to defect feature types.

4. External Corrosion Prediction

In this section, we take a high-level approach by producing machine learning (ML) architectures, which might be used to predict external corrosion or high-risk areas. We construct an ILI inspection dataset enriched with the geospatial features as defined in section 2. This dataset comprises joint level inspection data of over 5500 ILI inspections across the USA, and we aggregate geospatial features to the joint level; this means that the geospatial properties surrounding each joint are averaged in some manner. For this study, we define two predictive classes for the model: the positive class (P), for which external corrosion anomalies are present at the joint level to greater than 10% of wall thickness, and a negative class (N), where external corrosion anomalies are not present to this extent. We chose 10% as the threshold, the typical reporting depth from many ILI tools.

We consider data within the USA for two reasons. First, we guarantee consistent infrastructure legislation within the dataset and high-quality inspection data. Second, the availability and quality of the geo-enrichment features are of a high standard by exploiting the plethora of USA government surveys and data sources. The first point is important as we eliminate the possibility of spurious predictive power from the ML models which might be introduced from external variables not directly represented in the dataset but to which they are indirectly sensitive.

We perform a typical set of ML data preparation practices. The numerical data is normalized, and categorical data is one hot encoded. By standardizing in this way, we ensure that any one machine learning model is initially as sensitive to any one specific feature as any other, removing potential biases and improving model stability. We train two XGBoost [21] models, one with pipeline features only and one with geo-enrichment features with a randomized 90% train and a 10% test split. For XGBoost, we may choose a set of model hyperparameters that precisely control the training and model behavior; we select the best parameters by performing a hyperparameter grid search, which explores the full parameter space and picks the best combination based on our chosen validation metric. XGBoost is an efficient and powerful architecture that combines a series of weak classifiers, producing powerful and more accurate predictions when collated. The resultant dataset is highly imbalanced, due to most pipeline joints being in the negative class, comprising 90% of the data. Thus, it is essential to consider other metrics than accuracy; in our case, we take the balanced accuracy as a predictive metric of the model, but we also consider additional metrics such as f1 score, precision, recall, and area under the precision-recall curve.

4.1 Transparent Machine Learning

To dissect the model input features and their relevant importance, we use SHapley Additive exPlanations (SHAP) values [22]. SHAP values provide a way to glean insight into ML architectures. For any individual data point, we can understand why it was classified into a particular output class and which features pushed it in that direction. By marginalizing these SHAP values, we get a measure of feature importance.

In Figure 6, we see that SHAP values demonstrate that the geo-enrichment features in ML models hold high and equal importance to many of the pipeline features but also that many of the geo-enrichment features hold predictive power a small amount each. The SHAP analysis tells us explicitly that each feature contributes a small part to our overall understanding of the data.

4.2 Results

The geo-enriched XGBoost model has a balanced accuracy score of 76%, indicating good predictive power. However, there are further points to consider. Introducing a large number of feature inputs to an ML architecture has the potential to introduce confounding variables, which means we may have stifled the predictive power of the model by unjustifiably increasing model complexity. Models with high complexity can be more susceptible to poor model generalizability and bias. Further, with a deeply imbalanced dataset, we want to ensure that the model correctly assigns a high proportion of positive classes to the positive predictive class and the negative to the negative predictive class. Thus, the distribution of classifications must be studied with a confusion matrix.

Figure 7 shows the geo-enriched model confusion matrix. P represents the positive class, external corrosion greater than 10% depth of wall thickness, and N the negative class, less than 10% depth. We see that a significant proportion of the positive class and negative classes are incorrectly assigned. We can capture this misclassification by determining the precision and recall metrics. The precision of this model is 19%, and the recall is 73%. There is a balance to achieve between these two metrics; from a safety perspective, recall is the most appropriate metric as we are interested in minimizing false negatives and ensuring that if corrosion is found, the model predicts it. However, from a

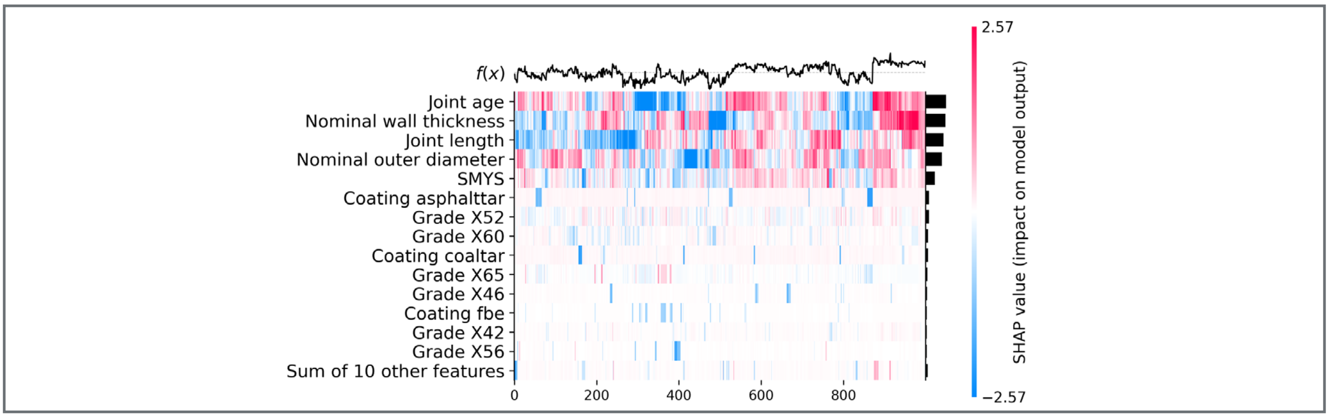


Figure 5: SHAP values for many datapoints (x-axis) represented in a heatmap for a model trained with only pipeline level features. The bars along the y-axis represent the sum of SHAP values for each feature which is a measure of feature importance ranking.

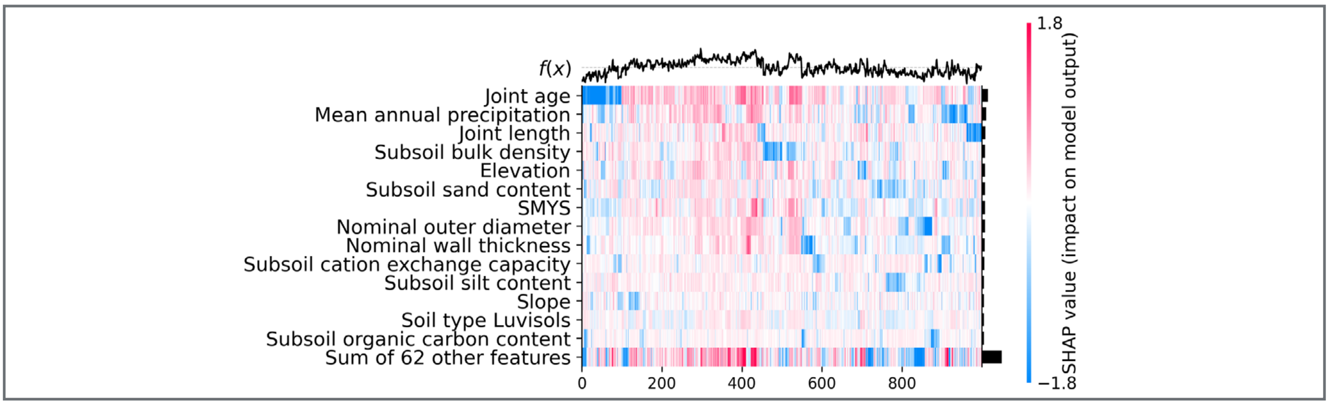


Figure 6: SHAP values for many datapoints (x-axis) represented in a heatmap for a model trained with pipeline level and geo-enrichment features. The bars along the y-axis represent the sum of SHAP values for each feature, which is a measure of feature importance ranking.

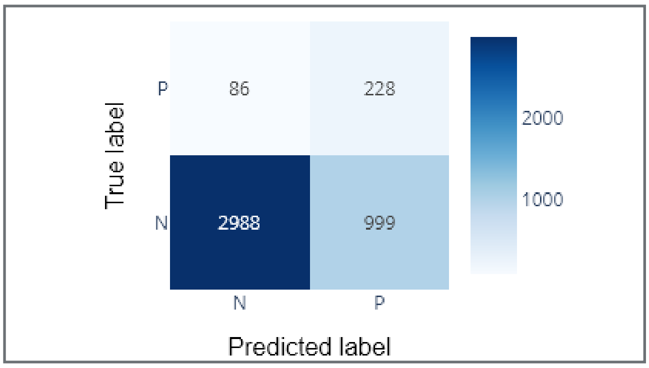


Figure 7: The geo-enriched model confusion matrix. Where P represents the positive class, external corrosion greater than 10% depth of wall thickness and N the negative class, less than 10% depth.

cost perspective, precision is more appropriate as we wish to minimize false positives such that erroneous recommendations for pipeline inspection are reduced. When developed further, models like this could be used as a decision support tool to advise engineers of regions of high risk, not provide precision predictions.

Further model validation and data engineering are required. Ongoing work is being performed to improve data aggregation methodology and the models capability to learn and make more meaningful predictions. Currently, the data aggregation and methodology are naïve; by predicting corrosion depth alone with a specific cut-off, we could omit shallow but large corrosion areas or clustering of defects. This methodology also makes assumptions on the aggregation of geospatial features; complex terrain details may be averaged out, which otherwise would be very powerful predictive features for our ML models. Slope units will provide a more natural way to aggregate geospatial features, and we have demonstrated that they may be used as a basis for predictive geohazard models.

The input data dimensionality should also be reduced, features with low feature importance should be removed and approaches such as Principal Component Analysis can reduce model complexity while retaining predictive power. Despite these short comings, these models illude to great potential in the dataset. We have gained insight into the most critical features and the validity of including expanded geospatial features in such models.

5. Outlook and Conclusions

This study demonstrates a relationship between slope units and their features and known areas of external corrosion or bending strain. However, this work is limited by the quantity of bending strain data; as more data is added to the IDW, there will be opportunities for deeper studies and the use of ML tools for which slope units could also be used directly as inputs. Slope units provide a natural basis for evaluating pipeline risk and generating natural aggregations of geospatial features, reinforcing geospatial relationships in the data. Additionally, new risk factors may be exposed by considering the spatial relationship between slope units and the sections of pipeline they contain; for example, the pipeline trajectory through a slope unit has not been considered here. There is also an opportunity to utilize an increased level of detail of the bending strain data, where more insight might be possible by considering strain measurements along the length of the pipeline, not just for sections above the reporting threshold.

Early results using ML models demonstrate clear classifying power for external corrosion predictions from IDW data, although there is significant room for improvement. Currently, there is a significant aggregation of the data, which limits any architecture's predictive power. Each data point is at the joint level, which means that the severity of external corrosion across any joint area with an anomaly of at least 10% wall thickness is all lumped together in the positive predictive class. Further, some joints may run over regions of terrain with changing features that get aggregated away. Future models will reconsider this approach; regions of the pipeline will be segmented at fixed lengths and fixed geospatial features. Despite these shortcomings, using SHAP values to assess feature importance, we discover that geo-enrichment feature inputs have comparable feature importance to pipeline property features when included in XGBoost models.

In response to the increasing challenge of geohazards driven by climate change, our research demonstrates the importance of analyzing and assessing combined integrity threats and coincident anomalies, indicating how ML techniques can be utilized to understand them. The work presented in this paper demonstrates the complex challenges of modeling external corrosion from ILI inspection data. The developed models show predictive power despite naïve approaches to data aggregation, and future work will aim to improve this methodology to fully exploit a range of geospatial features and develop more powerful predictive models.

References

1. M. Smith, L. Barton, K. Pesinis and I. Laing, "Intelligent Corrosion Prediction Using Bayesian Networks," in NACE CORROSION, Nashville, 2019.
2. M. Smith, S. Cronjaeger, N. Ershad, R. Nickle and M. Peussner, "Pipeline Data Analytics: Enhanced Corrosion Growth Assessment Through Machine Learning," in 12th International Pipeline Conference, Calgary, 2018.
3. M. Smith, A. Blenkinsop, M. Capewell and B. Kerrigan, "Now You SCC Me, Now You Don't: Using Machine Learning to Find Stress Corrosion Cracking," in 13th International Pipeline Conference, 2020.
4. J. D. Horton, C. A. San Juan and D. B. Stoeser, The State Geologic Map Compilation (SGMC) Geodatabase of the Conterminous United States, USGS, 2017.
5. L. Poggio, L. M. de Sousa, N. H. Batjes, G. B. M. Heuvelink, B. Kempen and E. Ribeiro, SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, vol. 7, SOIL, 2021, pp. 217-240.
6. M. E. Turek, L. Poggio, N. H. Batjes, R. A. Armindo, Q. de Jong van Lier, L. de Sousa and G. B. M. Heuvelink, Global mapping of volumetric water retention at 100, 330 and 15 000 cm suction using the WoSIS database, vol. 11, International Soil and Water Conservation Research, 2023, pp. 225-239.
7. U.S. Geological Survey, National Elevation Dataset, LIDAR Point Cloud (LPC), Groundwater Levels Service.
8. H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee and J.-N. Thépaut, ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2023.
9. U. G. Survey, "3D Elevation Program," [Online]. Available: https://www.usgs.gov/3d-elevation-program. [Accessed 02 10 2024].
10. U. G. Survey, "RockyWeb," [Online]. Available: https://rockyweb.usgs.gov/. [Accessed 02 10 2024].
11. J. Iwahashi and D. Yamazaki, Global polygons for terrain classification divided into uniform slopes and basins, vol. 9, Prog Earth Planet Sci, 2022.
12. J. Iwahashi and D. Yamazaki, Terrain22, 2022.
13. C. S. Juang, T. A. Stanley and D. B. Kirschbaum, Using citizen science to expand the global map of landslides: Introducing the Cooperative Open Online Landslide Repository (COOLR), vol. 14, PLOS ONE, 2019.
14. D. B. Kirschbaum, T. Stanley and Y. Zhou, "Spatial and temporal analysis of a global landslide catalog," Geomorphology, vol. 249, pp. 4-15, 2015.
15. D. B. Kirschbaum, R. Adler, Y. Hong, S. Hill and A. Lerner-Lam, "A global landslide catalog for hazard applications: method, results, and limitations," Natural Hazards, vol. 52, pp. 561-575, 2010.
16. Copernicus Climate Change Service (C3S), Climate Data Store (CDS), "Land cover classification gridded maps from 1992 to present derived from satellite observation," 2019.
17. OpenStreetMap contributors, Planet dump retrieved from https://planet.osm.org, 2017.
18. Z. Chang, F. Catani, F. Huang, G. Liu, S. R. Menna, J. Huang and C. Zhou, "Landslide susceptibility prediction using slope unit-based machine learning models considering the heterogeneity of conditioning factors," Rock Mechanics and Geotechnical Engineering, vol. 15, no. 5, pp. 1127-1143, 2023.
19. J. B. Woodard, B. B. Mirus, N. J. Wood, K. E. Allstadt, B. A. Leshchinsky and M. M. Crawford, Slope Unit Maker (SUMak): an efficient and parameter-free algorithm for delineating slope units to improve landslide modeling, vol. 24, Natural Hazards and Earth System Sciences, 2024, pp. 1-12.
20. "Celery - Distributed Task Queue," [Online]. Available: https://github.com/celery/celery.
21. Chen, Tianqi, Guestrin and Carlos, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
22. S. M. Lundberg and S.-I. Lee, A Unified Approach to Interpreting Model Predictions, Neural Information Processing Systems, 2017.
23. R. W. Saaty, The analytic hierarchy process—what it is and how it is used, vol. 9, Mathematical Modelling, 1987, pp. 161-176.
24. American Petroleum Institute, "API 579-1/ASME FFS-1," American Petroleum Institute, 2021.

AUTHORS

Joseph Walker
ROSEN Group
Data Scientist
jwalker@rosen-group.com

Edmund Bennett
ROSEN Group
Principal Data Scientist
ebennett@rosen-group.com

Jonathan Martin
ROSEN Group
Senior Data Engineer
jmartin@rosen-group.com

Jozef Soltis
ROSEN Group
Principal Corrosion Engineer
jsoltis@rosen-group.com