**Introduction**

As the global energy landscape evolves, the petroleum industry continues to grapple with the challenges of ensuring environmental safety and human well-being. This introductory article lays the foundation for an extensive exploration of flowline risk analysis within the oil and gas sector. Our ultimate goal is to mitigate environmental impacts and prevent human casualties associated with flowline failures through risk analysis using GIS and machine learning methods. However, a critical hurdle in this endeavor is the current lack of adequate data necessary to conduct such analysis. Therefore, this paper serves to provide a background on flowlines, review existing risk analysis literature, summarise our work with the data at hand along with its limitations, and outline future expected work after receiving the necessary data.

Flowlines, often overshadowed by the more visible pipelines, are crucial components in the petroleum production process. These underground conduits transport oil, natural gas, and water from wellheads to surface facilities and ultimately to Lease Automatic Custody Transfer (LACT) units. This research focuses on the 'middle half' of the oil and gas production process, stretching from wellheads to LACT units, where most U.S. flowlines are buried to prevent freezing and maintain structural integrity. Understanding the nuances of flowlines, including their material standards set by the American Petroleum Institute (API), construction, operational dynamics, and reasons for failure, is critical for a comprehensive risk assessment.

Building on this understanding, the innovative approach of this research lies in the application of machine learning and GIS in assessing and mitigating risks associated with flowlines. Machine learning algorithms are adept at identifying patterns and predicting potential failures by analyzing vast datasets, which traditional methods might overlook. Coupled with GIS's spatial analysis capabilities, this approach provides a greater understanding of the geographical factors affecting flowline integrity. This integrated method is expected to yield a more accurate and comprehensive risk assessment model, enhancing the safety and reliability of the petroleum extraction process. However, a notable challenge that emerges in this process is the significant gap in our knowledge regarding flowline failures. This deficiency severely limits our ability to conduct an in-depth analysis that can predict and model the likelihood and consequences of flowline failures.

This introduction sets the stage for a detailed exposition of our research and data needs. We aim to provide a clear and comprehensive view of the complexities involved in flowline risk analysis and the transformative potential of machine learning and GIS in this field. Through this paper, we seek to engage our industry partners in a collaborative effort to gather the necessary data, paving the way for significant advancements in the safety and environmental stewardship of petroleum operations.

**Literature Review**

Prior to data exploration, the team conducted an extensive literature review to understand existing efforts in flowline/pipeline risk assessment. Numerous national and international organizations have developed models using various data types to predict the risks associated with flowlines/pipelines. Their main objective is to service or replace high-risk lines before they result in spills. Spills not only incur substantial costs for the responsible organization but also cause significant environmental damage and can lead to loss of human life, especially when they occur near residential areas. This has prompted many organizations to seek solutions to this challenge.

Risk models have been developed primarily using two approaches: mathematical computation models and machine learning models. One mathematical model applied regression to real-life data, calculating influence coefficients for each data input based on actual pipeline failures. It also assessed the potential damage of each pipeline's failure to prioritize high-risk pipelines (Vinogradov et al., 2018). In contrast, various machine learning models have been employed which are primarily data-driven rather than following theoretical justifications. For instance, one study compared three machine learning models—log-linear regression, eXtreme Gradient Boosting (xgBoost), and Artificial Neural Networks—to predict corrosion growth and found xgBoost to be the most accurate (Mazzella et al., 2019). Another study demonstrated the use of Euclidean-Support Vector Machines to predict pipeline failure using continuous sensor data (Lam Hong Lee et al., 2013).

Across the literature, a combination of available data was used to develop these models. The data's role is crucial in determining the reliability and accuracy of these risk models. The main categories of data used across the publications include pipeline specifications (diameter, age, coating, etc.), GIS data, soil data (for buried flowlines), human activity data (such as proximity to roads), inspection reports, repair/service history, historical incident records, operational data (flow rate, type of transported fluid, pressure, etc.), and continuous monitoring sensors data (for example data from Long Range Ultrasonic Transducers) (Vinogradov et al., 2018), (Mazzella et al., 2019), (Zhang & Liu, 2023), (Khalilpasha & Brown, 2023), (Guan et al., 2019), (Senouci et al., 2014), & (Lam Hong Lee et al., 2013). Due to challenges in data availability and collection, some studies used simulated data as proof of concept (Lam Hong Lee et al., 2013)., while others employed various combinations of the mentioned data categories to develop their risk models.

The output of these publications largely depends on the data used to train the models. The literature primarily focuses on predicting corrosion, corrosion growth rate, risk ranking, remaining life, or type of failure. Flowline/pipeline failures typically stem from design issues, manufacturing issues, installation issues, corrosion and erosion, structural threats (fatigue, static overload, etc.), natural hazards, and human error (Rachman et al., 2021). Predicting such outcomes is crucial for minimizing spill incidents, as it enables operators to identify and address high-risk pipelines proactively and be one step ahead of flowlines/pipelines failures.

The effectiveness of a risk management model for flowlines/pipelines heavily depends on the accuracy, quantity, and diversity of the data used for training. Reliable data significantly enhances the model's predictive capabilities. However, it's important to acknowledge that the model's capacity is limited to the data it's trained on. For instance, a model trained on pipeline specifications, GIS data, and inspection reports might be highly accurate, but if it lacks training on human activity and operational data, it may miss failures linked to these factors. Therefore, the careful selection of input data is a key factor in the model's success. When choosing data for the model, the following should be considered:

* The relevance of features to failure determines the model's effectiveness.
* The cost of data acquisition; unsustainable data collection can lead to project failure.
* The reliability of the data; unreliable data results in an unreliable model.

The sufficiency of data to ensure a comprehensive coverage of the entire area of interest.

**Table 1 summarizes the input data, methodology, model output, and feature availability in various literature publications:**

**Table-1:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Category | Feature | Methodology | Output | Resource | Provided by ECMC? |
| Pipeline Specification | Service Life (age) | Regression + Coefficients of influence assigned to each parameter | Pipelines risk ranking | (Vinogradov et al., 2018) | Could be estimated using wells age |
| Pipeline material | Yes |
| Inhibition | No |
| Presence of Internal protection | No |
| Presence of external protection | No |
| Avg annual corrosion rate | No |
| Operational Data | Pipeline purpose | No |
| Water content in transported product | No |
| H2S in transported product From wells | No |
| Flow rate | No |
| Environmental | Soil corrosion activity Potentially from USDA soil survey | No |
| Presence of stray currents | No |
| Historical incident records | Number of failures during latest year of operation  (see spills records) | Maybe |
| Total number of failures  (see spills records) | Yes |
| Inspection reports | Employed diagnostic methods | No |
| Repair/service history | In-line cleaning | No |
| Environmental Soils data Potentially from USDA soil survey | SO2 per m2/d (Atmosphere) | Machine Learning Models:  1.Log Linear Regression  2.xgBoost  3.Artificial Neural Networks  90% Training  10% Validation | Corrosion Growth Rate | (Mazzella et al., 2019) | No |
| Cl per m2/d (Atmosphere) | No |
| Avg time of wetness (Atmosphere) | No |
| Annual Avg Temp. (Atmosphere) | Yes |
| Avg days below 0 C  (Any dip > 0 C?) | Maybe |
| Over Land (0) or Water (1) | No |
| Organic Carbon (Soil) | No |
| pH | No |
| Silt | No |
| Sand, mass% (Soil) | No |
| Clay, mass% (Soil) | No |
| Water Content | No |
| Electrical Conductivity (Soil) | No |
| # of nearby electric stations | No |
| # of AC lines withing 300m | No |
| Magnetic Annamolly Value | No |
| Line is within 300m of > 300 voltage | No |
| Powerline within 100 m | No |
| AC substation within 500m | No |
| Nearby powerline > 100 voltage | No |
| Human Activity  County DOT  [(Weld)](https://gishub.weldgov.com/datasets/4d5e17afe57f474fb28baee35e942031_0/about) | Roads within 100 m with Max speed  > 40 MPH | No |
| Roads within 100 m with Avg speed > 40 MPH | No |
| # of roads within 100 m that > 40 MPH | No |
| Nearby operational railway  [(DOT Railway Data)](https://geodata.bts.gov/search?tags=Rail) | No |
| # of nearby pipelines | No |
| Pipeline Specifications | Manufacturer | No |
| Year of Mill Run | No |
| Actual outer diameter  [(Can be implied w/ nominal vs. actual data)](https://www.engineersedge.com/pipe_schedules.htm) | No |
| Pipeline Specifications | | Multiple Linear Regression & Advanced Machine Learning Regression Methods | Leak Time | (Zhang & Liu, 2023) | No |
| GIS Data | | Yes |
| Inspection Data | | Cognitive Learning (ML) | Remaining Pipeline Life  (Carries Gas & Condensate in Australia) | (Khalilpasha & Brown, 2023) | Yes |
| GIS Data | | No |
| Pipeline Specifications (Including Cathodic Protection) | | No |
| Inspection Reports | | Yes |
| Repair/service history | | No |
| Pipeline Specifications | Coating type | Bayesian Networks | External& Internal Corrosion Risk | (Guan et al., 2019) | No |
| Cathodic protection and effective surface potential | No |
| Operational Data | Pressure | No Some |
| Sulfates content | No See well data |
| Chlorides content | No See well data |
| Environmental | [Topography](https://edcftp.cr.usgs.gov/Contours/preliminary_by_state/CO_preliminary.gdb.zip) | Maybe |
| Soil Type  [USGS Web Soil Survey](https://websoilsurvey.nrcs.usda.gov/app/) | No |
| Excavation Data | Maybe |
| Crossing locations | No Xref w/DOT data and large water bodies |
| Stray currents | No |
| Direct Current Voltage Gradient | No |
| Locations of dents, welds, & bends | No Welds could be implied with newer data |
| Operational Data | Type of transferred product | Regression & Artificial Neural Networks | Failure Type:  1.Corrosion  2.Natural Hazard  3.Mechanical  4. Operational5.Third Party | (Senouci et al., 2014) |  |
| Environmental Data | Land use | Yes |
| GIS Data | Pipe Location | Yes |
| Pipe Specification | Pipe Age | No See wells spud date |
| Pipe Diameter | Yes |
| Continuous Monitoring Sensors | Long Range Ultrasonic Transducers | Euclidian support vector machine learning | Corrosion depth | (Lam Hong Lee et al., 2013) | No |

**Work Done**

Commissioned by the Colorado Energy and Carbon Emission (ECMC), our research team has been tasked with designing a flowline risk assessment model. Contrary to previous qualitative risk assessment models, this project takes a more quantitative approach by incorporating modern machine learning techniques. To better comprehend the problem, we deconstruct risk assessment into two fundamental components: the probability and consequence of potential pipeline hazards. These two criteria, in combination, create a risk matrix, which assigns numerical risk levels to given pipelines.

The initial step of our work was to ensure the data available reflects the problem at hand. In detail, understanding the data representation and accessing the integrity of the datasets became crucial. With these concerns, our team explored the various publicly available datasets on the ECMC website, including but not limited to the Mechanical Integrity, Spills, Field Inspection, and GIS datasets. As the topic relates to risk assessment, our team meticulously examined the Spills dataset. As of November 2023, the Spills dataset comprises over 15428 observations and 106 columns. While the dataset boasts substantial features, the bulk of the Spills dataset remains incomplete or missing.

For context, the Spills dataset consists of two(three) report forms: the initial report forms and the supplementary report forms (along with initial report forms containing supplementary materials). The initial report forms contain identifying information about the incident, such as document number, operator name, tracking number, etc., as well as estimates of spilled fluid volumes and types. The initial report forms account for approximately 40 out of the 106 features in the Spills dataset and are mostly complete. However, most of these features in the initial report form are not pivotal for machine learning purposes; instead, the saliency lies in facilitating record-keeping for database management.

The supplementary forms, thus, populate the remaining columns of the dataset, consisting of geospatial features, failure classifications, root causes, and more. Due to the data collection and reporting process, a pipeline incident incorporates multiple supplementary forms or lacks the associated supporting materials. Consequently, by merging corresponding documents, the original 15,428 entries of the dataset can be condensed to approximately 6,945 unique pipeline spill incidents.

Upon closer scrutiny of these entries, focusing on the supplementary form materials in the context of the risk modeling process, it becomes apparent that deficiencies persist in the dataset. While the supplementary materials offer more pertinent features to the machine learning process, its presence in the dataset is notably sparse. These data gaps manifest in two ways: either a substantial portion of observations in the dataset does not have a corresponding supplementary form, or the forms are not completely populated. Incomplete incident reports pose challenges, particularly in assessing the consequences of related pipeline spills, as the requisite information is not readily available.

Furthermore, despite encompassing over 106 features, the Spills dataset lacks elements crucial for quantifying the likelihood of a pipeline hazard. Currently, the sole indicator is derived from the 'root cause' column, a text box entry describing the cause of a pipeline spill. Although feature extraction from these text descriptions is possible, the results would predominantly function as outputs for our classification problem, categorizing the rationale behind pipeline incidents. However, this does not alleviate the issue of limited feature representation and predictive powers within the Spills dataset.

Also, we observe that the data collection process is reactive, not concurrent. Many of the details regarding the spill are reported later than when the spill initially occurred. In essence, the Spills dataset alone offers limited possibilities for analysis or intervention.

While delving into the dataset, our team conducted exploratory data analysis to capture insights about the dataset. The findings are succinctly presented in the following dashboard.

A screenshot of a computer

Description automatically generated

A few things to note regarding the dashboard:

* Aggregated functions such as the mean are employed to represent geospatial data, primarily to address the significant amount of missing data observed in the Spills dataset
* The breakdown of spill volumes highlights the extensive amount of data that is aggregated under the categories of unknown or NA
* This dashboard was in use until November 2023, with the dataset's last update recorded in October of the same year. Access to the dashboard can be found [here](https://public.tableau.com/app/profile/zhang.lin2425/viz/ECMC_Dashboard/PipelineSummary)

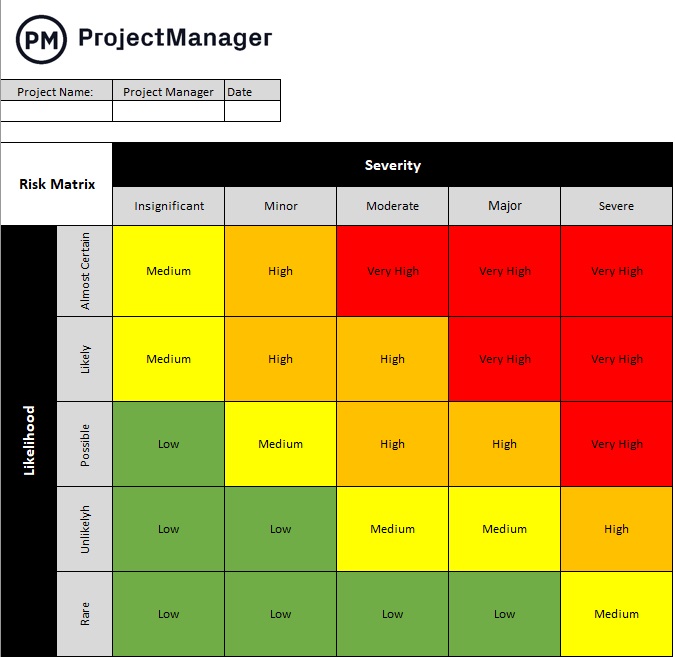
In our pursuit of completing the data required for the project, our team explored various datasets and databases within ECMC. Despite our efforts, we were unable to find the necessary information. The quest for predictive features, particularly operational parameters, led us to investigate the Mechanical and Inspection datasets. Unfortunately, the datasets, though containing some operational parameters, proved largely incomplete. Additionally, querying the correct information across different databases posed significant challenges, leaving us uncertain about the databases' design to support our actions.

From a database perspective, each spill incident is linked to a unique tracking number, and document numbers serve as the unique keys to each database/form. This limitation eliminates the possibility of unique joins, restricting us to operator or facility identification numbers that are not supported in particular databases. Attempts at joins opened up the potential for many-to-many joins, where the distinction of each spill incident couldn't be established. Our initial attempts at data consolidation yielded limited and unsuccessful results, prompting the need for further data acquisition.

In an effort to enhance geospatial information for consequence analysis, we turned our attention to GIS data. However, public, and private spatial files merely provided the anticipated spatial representations of Colorado pipelines. While we acknowledge the potential insights this information could offer, as of now, the research team has not taken further action.

**Future Work**

**­­**Once we have cleaned our data and roughly identified relevant predictive features, we will begin the modeling phase of the project. Each model we create will calculate the probability of failure (POF) and consequence of failure (COF) for each flowline. POF represents the likelihood of an accident occurring, and COF represents the impact of such an incident, including financial loss, environmental damage, and impact on nearby human activity. Overall risk is the product of POF and COF, which can be represented by a risk matrix. An example risk matrix is shown below.



We expect that to accurately model POF and COF, we will need to use a variety of inputs. Pipeline data such as pressure, temperature, flow rate, material, fluid transported, diameter, wall thickness, etc., will likely prove critical to quantifying POF. To calculate COF, we will also need to incorporate information about the surrounding area, including nearby residences, wells, etc. We plan on determining these characteristics by combining the private GIS data set with other available data sets that contain the relevant information. Our list of relevant features will be refined as we continue to explore the data we already have and learn what other data is available.

We plan on assessing the quality of our models using multiple evaluation metrics. The most intuitive metric is accuracy:

Accuracy is useful for evaluating predictive power of a model because it is simple and easy to interpret, but it lacks the specificity of some slightly more sophisticated metrics. To augment accuracy, we will create a confusion matrix, which distinguishes between true positives, true negatives, false positives, and false negatives. We will also calculate precision and recall as follows:

We will begin modeling with a simple multiple linear regression model which follows the form:

Where is a constant term, are the input features, are the corresponding weights for the independent variables, is the residual error term, and is the output. We will create separate linear regression models for COF and POF, with different variables and individual tuning of the weights. The overall risk for a pipeline will then be the product of the outputs of these two models. Depending on the characteristics of the data, it may be appropriate to perform a logarithmic transformation on some variables to create more accurate models.

Our initial linear model will likely not demonstrate satisfactory predictive power, but it will provide us with useful insight regarding features, and the relationships between features and our desired outputs. Based on what we learn from the linear model, we will move forward with more sophisticated modeling techniques such as neural networks or decision trees.

**References**

* Vinogradov, P. V., Litvinenko, K. V., Valiakhmetov, R. I., and A. N. Bakhtegareeva. "Development of a model for ranking field pipelines based on risk assessment in exploitation (Russian)." OIJ 2018 (2018): 84–86. doi: https://doi.org/10.24887/0028-2448-2018-8-84-86
* Mazzella, Joseph, Hayden, Thomas, Krissa, Len, and Haralampos Tsaprailis. "Estimating Corrosion Growth Rate for Underground Pipeline: A Machine Learning Based Approach." Paper presented at the CORROSION 2019, Nashville, Tennessee, USA, March 2019.
* Zhang, Zhuoran, and Guanlan Liu. "A Prediction of Corrosion-Related Leakage on Distribution Pipelines via Machine Learning Method." Paper presented at the AMPP Annual Conference + Expo, Denver, Colorado, USA, March 2023.
* Khalilpasha, Hossein, and Justin Brown. "Minimizing Pipeline Leaks and Maximizing Operational Life by Application of Machine Learning at Cooper Basin." Paper presented at the AMPP Annual Conference + Expo, Denver, Colorado, USA, March 2023.
* Guan, Shan, Ayello, Francois, Sridhar, Narasi, Han, Xiaoming, Feng, Qingshan, and Yonghe Yang. "Application of Probabilistic Model in Pipeline Direct Assessment." Paper presented at the CORROSION 2019, Nashville, Tennessee, USA, March 2019.
* Ahmed Senouci, Mohamed Elabbasy, Emad Elwakil, Bassem Abdrabou & Tarek Zayed (2014) A model for predicting failure of oil pipelines, Structure, and Infrastructure Engineering, 10:3, 375-387, DOI: 10.1080/15732479.2012.756918
* Lam Hong Lee, Rajprasad Rajkumar, Lai Hung Lo, Chin Heng Wan, Dino Isa. Oil and gas pipeline failure prediction system using long range ultrasonic transducers and Euclidean-Support Vector Machines classification approach. Expert Systems with Applications, Volume 40, Issue 6, 2013, Pages 1925-1934, ISSN 0957-4174. <https://doi.org/10.1016/j.eswa.2012.10.006>.
* Andika Rachman, Tieling Zhang, R.M. Chandima Ratnayake, Applications of machine learning in pipeline integrity management: A state-of-the-art review, International Journal of Pressure Vessels and Piping, Volume 193, 2021, 104471, ISSN 0308-0161, https://doi.org/10.1016/j.ijpvp.2021.104471.