Risk Analysis of Flowlines in the Oil and Gas Sector: A GIS and Machine Learning Approach

I. Chittumuri[1], N. Alshehab[2], R. J. Voss[1], L. L. Douglass[1], S. Kamrava[2], Y. Fan[2], J. Miskimins[2], W. Fleckenstein[3], S. Wheeler[4], and **S. Bandyopadhyay[1*]**

[1]Department of Applied Mathematics and Statistics, Colorado School of Mines
[2]Department of Petroleum Engineering, Colorado School of Mines
[3]Office of Global Initiatives and Business Development, Colorado School of Mines
[4]Colorado Energy and Carbon Management Commission

*Corresponding author; email: `sbandyopadhyay@mines.edu`

Keywords: Risk Analysis; Flowlines; Machine Learning; GIS; Principal Component Analysis; Canonical Correlation Analysis;

Summary

This paper presented a risk analysis of flowlines in the oil and gas sector using Geographic Information Systems (GIS) and Machine Learning (ML). Flowlines, vital conduits that transported oil, gas, and water from wellheads to surface facilities, often faced under-assessment compared to transmission pipelines. This study addressed this gap using advanced tools to predict and mitigate failures, improving environmental safety and reducing human exposure. Extensive datasets from the Colorado Energy and Carbon Management Commission (ECMC) were processed through spatial matching, feature engineering, and geometric extraction to build robust predictive models. Various ML algorithms, including logistic regression, support vector machines, and gradient boosting decision trees, were used to assess and classify risks. To explore the impact of feature space reduction, two dimensionality reduction methods were applied: Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA). Finally, a thorough data analysis highlighted spatial and operational factors that influenced risks, identifying high-risk zones for focused monitoring. Overall, the study demonstrated the transformative potential of integrating GIS and ML in flowline risk management, proposing a data-driven approach that emphasized the need for accurate data and refined models to improve safety in petroleum extraction.

Introduction

As the global energy landscape evolved, the petroleum industry continued to face challenges of ensuring environmental safety and human well-being. In this work, we laid the foundation for an extensive exploration of risk analysis of flowlines within the oil and gas sector. Our ultimate goal was to mitigate environmental impacts, improve safety, and reduce human exposure associated with flowline failures through risk analysis using GIS, ML, and dimensionality reduction. However, a critical hurdle in this endeavor was the lack of adequate data necessary to perform such an analysis. Therefore, this paper served to provide background on flowlines, review the existing literature on risk analysis, summarize how we approached cleaning and analyzing the data at hand, and outline possibilities for future work.

Flowlines, often overshadowed by more well-studied 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 focused on the 'middle half' of the production process, where most U.S. flowlines were buried to prevent freezing and maintain structural integrity. Understanding the nuances of flowlines, including their material standards set by the American Petroleum Institute, construction, operational dynamics, and reasons for failure, was critical for a comprehensive risk assessment.

While previous studies have explored GIS or ML methods for pipeline risk analysis, they often treated spatial data simplistically or relied on simulated inputs. Our study presented a novel integration of multilinestring GeoDataFrames—a complex spatial data structure—into a full machine learning workflow. This allowed us to extract geometric descriptors (e.g., line complexity, spatial footprint) and convert them into model-ready features, improving interpretability and predictive power. Additionally,

we implemented dimensionality reduction techniques to evaluate their effectiveness in managing high-dimensional feature spaces. These techniques helped mitigate overfitting and improve computational efficiency while preserving predictive accuracy.

Our work centered specifically on flowlines—a segment of petroleum infrastructure often underrepresented in existing research—using real-world regulatory datasets from the Colorado Energy and Carbon Management Commission (ECMC). To our knowledge, no previous study had developed a replicable and open-source ML-GIS pipeline focused on this infrastructure class, making this effort a step forward in practical and scalable environmental risk assessment. ML algorithms in our study were adept at identifying patterns and predicting potential failures by analyzing large, complex datasets that traditional methods might have overlooked. By extracting key spatial features from multilinestring geometries and incorporating them into ML models, we enabled the algorithms to utilize geometric and locational data more effectively. This integrated method ultimately yielded a precise risk assessment model, enhancing the safety and reliability of petroleum extraction processes.

However, a significant challenge in this effort was the lack of comprehensive data on flowline failures, which limited the depth of analysis and the ability to fully predict the likelihood and consequences of such failures. This introduction sets the stage for a detailed exposition of our research and data needs. Through this work, we aimed to provide a clear and comprehensive view of the complexities involved in flowline risk analysis and to highlight the transformative potential of combining ML and GIS. We also sought to engage our industry partners in a collaborative effort to expand data access, paving the way for further advancements in the safety and environmental stewardship of petroleum operations.

Before embarking on data exploration, we performed an extensive literature review to understand existing efforts in flowline and pipeline risk assessment. Several national and international organizations developed models using diverse data types to predict risks associated with flowlines and pipelines. The primary goal of these models was to service or replace high-risk lines before they led to spills, which could cause significant environmental damage and potentially result in loss of human life, especially in densely populated areas. This need drove many organizations to seek out effective solutions to mitigate these risks.

Risk assessment models were primarily developed using two approaches: mathematical computation models and ML models. For instance, one mathematical model applied regression analysis 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). On the other hand, various ML models were employed, which were primarily data-driven rather than based on theoretical justifications. A study by Mazzella et al. (2019) compared three ML models—log-linear regression, eXtreme Gradient Boosting (xgBoost), and Artificial Neural Networks—to predict corrosion growth and found xgBoost to be the most accurate. Another study demonstrated the use of Euclidean-Support Vector Machines to predict pipeline failure using continuous sensor data (Lam Hong Lee et al., 2013).

The literature showed that a combination of available data was used to develop these models. The role of data was crucial in determining the reliability and accuracy of these risk models. The main categories of data used across the studies included 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 sensor 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; and Lam Hong Lee et al., 2013). Some studies employed simulated data as a proof of concept (Lam Hong Lee et al., 2013), while others used various combinations of the mentioned data categories to develop their risk models. The output of these models largely depended on the data used to train them. The literature primarily focused on predicting corrosion, corrosion growth rate, risk ranking, remaining life, or type of failure. Flowline and pipeline failures typically stemmed from design issues, manufacturing flaws, installation errors, corrosion and erosion, structural threats (such as fatigue and static overload), natural hazards, and human error (Rachman et al., 2021). Predicting such outcomes was crucial for minimizing spill incidents, as it enabled operators to proactively identify and address high-risk pipelines, staying one step ahead of potential failures.

The effectiveness of a risk management model for flowlines and pipelines heavily depended on the accuracy, quantity, and diversity of the data used for training. Reliable data significantly improved the model's predictive capabilities. However, it was important to acknowledge that the model's capacity was limited to the data it was trained on. For example, a model trained on pipeline specifications, GIS data, and inspection reports might have been highly accurate, but if it lacked training on human activity and

operational data, it could have missed failures linked to these factors. Therefore, careful selection of input data was a key factor in the model's success. When choosing data for the model, the relevance of features to failure determined the model's effectiveness, the cost of data acquisition needed to be sustainable, the reliability of the data had to ensure an accurate model, and the sufficiency of data had to cover the entire area of interest.

Data Description and Processing

The Colorado Energy and Carbon Management Commission (ECMC) maintain a range of datasets accessible on their website. Among these, the "Spill Data" file was particularly pertinent to our study, encompassing 1,726 initial individual cases for each spill recorded in Colorado from 2014 to 2022. These cases were documented with an extensive array of variables, including latitude, longitude, volume spilled, root cause, etc. For more specifics on this dataset, please see the ECMC website provided at the conclusion of this document. To augment the spills dataset, ECMC provided two additional flowline datasets. The first, referred to as the descriptive dataset, included detailed GIS data pinpointing the exact locations of entire flowlines, along with basic attributes such as fluid type, material, diameter, and operational status. The second, known as the operational dataset, contained only the start and end points of flowlines but provided important operational information essential for constructing a robust risk model. This operational information included location type, maximum operating pressure, and construction date. In total, these two flowline datasets initially comprised of 21,942 flowlines.

The first step in our data processing was to amalgamate the descriptive and operational datasets into a unified flowline dataset. Although the datasets included identification parameters, such as Operator Names and Flowline IDs, these were not unique and could correspond to multiple flowlines, which limited their reliability for one-to-one spatial matching. Therefore, due to the absence of unique identifiers, spatial characteristics served as the basis for integration.

The operational data was converted into a GeoDataBase (.gbd) format, with GPS coordinates adjusted to the EPSG:26913 coordinate reference system used in the descriptive data. We approximated each flowline by interpolating straight lines between the provided endpoints and merged the two datasets by identifying intersections at both endpoints. Given the spatial imprecision in the data and ECMC industry standards, we implemented a tolerance mechanism, beginning at 0 meters and incrementally expanding up to 25 meters. If no match was found within this range, the line was excluded.

Fig. 1 below illustrated this challenge. It depicted an approximated flowline (in red) from the operational data overlapping multiple precise flowlines (in blue) from the descriptive data. The lack of exact endpoint alignment complicated each match. To minimize errors, operator names were used as an additional verification step; any mismatch in operator names, even within the tolerance range, led to the continuation of the search for a more accurate match. This, however, did not entirely eliminate inaccuracies, as many of the flowlines originating from the same facility were often maintained by the same operator, so overlaps in operator names across proximal flowlines were common. By applying the matching approach and removing rows with missing data, we approximately linked attributes from the operational dataset to the descriptive data, resulting in only 4,117 matched flowlines. It should be noted that mismatches likely occurred due to the high density of flowlines and inherent inaccuracies within the spatial data.

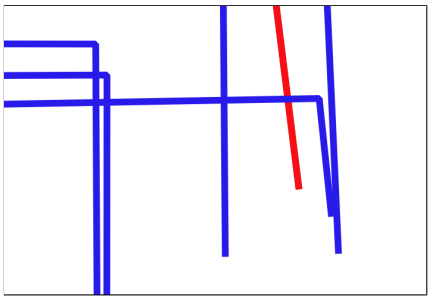

Fig. 1: Spatial Matching Challenge in Flowline Data Integration

Subsequent to establishing a cohesive flowline dataset, we attempted to correlate each spill to its originating flowline. First, we removed spills associated with gathering lines from the dataset. While these lines were present in the spill reports due to ECMC's role in overseeing spill reporting and remediation,

they were not within ECMC's regulatory jurisdiction, as gathering line safety was governed by the Public Utility Commission's Pipeline Safety Program. Given that this study was developed in response to ECMC's request to assess risk for flowlines specifically, we excluded gathering line spills to align with the regulatory scope. After filtering, the dataset was reduced from 1,726 to 849 spill entries, each represented as a spatial point. Then, we employed a spatial matching process with an increasing tolerance from 0 to 25 meters to find the first intersecting flowline, using operator names for validation. However, the single-point comparison significantly heightened the risk of incorrect associations, as illustrated in **Fig. 2**.

Fig. 2 depicted green points (spills) near blue lines (flowlines). The methodology assumed the nearest flowline with a matching operator name as the source, but this assumption was vulnerable to errors in spill location reporting. Often, a spill occurred underground and then moved laterally through the Earth before reaching the surface, where its location was recorded. In conjunction with measurement inaccuracies, this meant it was plausible that the flowline from which a spill originated was not necessarily the closest flowline to the recorded location of the spill. Thus, a spill might have been incorrectly matched to a nearby line rather than its true origin. After using this method and removing any rows without data, we ended up with only 41 spills approximately matched to a flowline.

The 25-meter matching threshold was chosen based on guidance from our partners at ECMC, who emphasized the importance of minimizing false positives in spatial joins. Given the potential for multiple overlapping flowlines at the same site, a more generous tolerance risked incorrectly linking spills to unrelated infrastructure. While this conservative approach improved confidence in matches, it also significantly reduced the usable dataset to approximately 20%. This limitation reflected constraints in the historical geospatial recording of flowlines and spills, rather than a shortcoming of the modeling process.

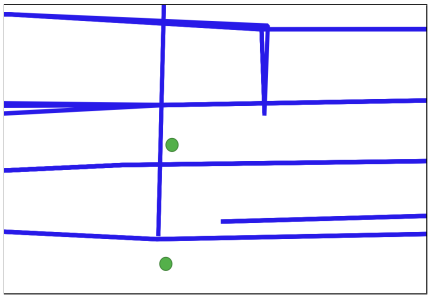

Fig. 2: Illustration of Potential Misalignment in Spill Matching

Feature Engineering

Next, we concentrated on feature selection and engineering to refine our datasets for more effective analysis. One key attribute we developed was the age of each flowline, which we inferred from the construction date. Understanding the age of flowlines was vital, as age might have correlated with spill likelihood. Specifically, older flowlines were hypothesized to be more prone to failures due to degradation over time. To differentiate flowlines based on their spill history, we introduced a binary risk column. This column labeled each flowline as either high risk (1) if associated with a spill, or low risk (0) if no spills were recorded. However, this label should be interpreted as a soft indicator, as the absence of recorded spills did not necessarily imply low risk—some spill events may have been unreported or undocumented. Thus, while the risk column provided a useful starting point, our modeling framework treated it as a proxy for underlying risk, allowing us to explore which flowlines without recorded spills exhibited similar characteristics to those with confirmed incidents.

The features used in our model were selected based on three key criteria: (1) their theoretical relevance to pipeline failure mechanisms (e.g., age, material, pressure); (2) support in prior risk modeling literature; and (3) guidance from ECMC subject matter experts, based on data availability and known risk factors. As part of our feature selection, we excluded identifiers such as Operator Number, Flowline ID, and Location ID from the modeling dataset. These fields functioned primarily as non-unique identification parameters rather than informative predictors and could introduce noise or overfitting, especially when one-hot encoding them due to their high cardinality. While they were used in the exploratory data

analysis (e.g., operator-level trends), they were dropped prior to model training. However, since many attributes in our dataset were categorical and lacked inherent ordering (nominal data), one-hot encoding was required. One-hot encoding converted each category value into a new binary column, ensuring that the model interpreted these variables correctly without assuming a natural ordering. Attributes such as Status, Flowline Action, Location Type, Fluid Type, and Material were transformed. This technique prevented the introduction of arbitrary ordinality that numeric encoding might have suggested.

One key variable included in our dataset was root cause type, which was also one-hot encoded to identify the reported cause of a spill when available. Categories included Corrosion, Incorrect Operation, Natural Force Damage, Other Outside Force Damage, Equipment Failure, and Unknown. This variable was only present for flowlines with recorded spill events; all others were assigned NaN for this field. Given the limited number of flowlines with known root causes and the absence of many parameters traditionally associated with failure modeling (e.g., soil corrosivity, inspection history, pressure cycling), our approach focused on generalized risk classification rather than failure-mode-specific prediction. While our feature selection focused on variables with known relevance and consistent availability, we acknowledged that some excluded attributes (e.g., administrative fields, incomplete categorical labels) might have held predictive value if more complete or context-specific information had been available.

Extracting meaningful features from the Geometry column, which was the geo-spatial location data of the flowline in our dataset, was crucial since ML models could not directly interpret geometric data. We decided to tabularize the data, extracting the most descriptive features to use in our model. This was the simplest and most intuitive approach to using GIS data as inputs for ML models. We derived two key features: Number of Lines and Bounding Box Area. The Number of Lines feature quantified the complexity of the geometry by counting the number of line segments it contained, differentiating between simple and complex structures. The Bounding Box Area was determined by calculating the area of the smallest rectangle that could completely enclose the geometry, providing insights into the spatial extent of each flowline. These features transformed the geometric data into quantitative values that could be readily used in predictive modeling, improving our ability to assess and manage risks associated with the flowlines. By transforming complex spatial data into these features, the dataset became more actionable for predictive modeling. These derived metrics provided insights into the physical characteristics and spatial distribution of the flowlines, which were instrumental in predicting and mitigating spill risks effectively.

Another significant challenge in our data processing was addressing inconsistencies in the dataset, such as variations in spelling and formatting across various fields. Standardizing these elements was crucial for maintaining data integrity and ensuring accurate analysis. We tackled these discrepancies by mapping misspelled or inconsistently formatted variables to their correct versions. This mapping process was applied to multiple columns, ensuring that all entries were uniformly formatted. This step was essential to prevent misclassification and errors during data processing, making the dataset cohesive and reliable for advanced statistical analyses and ML applications. In addition, to ensure data integrity and avoid data leakage, each flowline was treated as a unique sample identified by Flowline ID, and duplicate entries were removed during preprocessing. The final dataset consisted of exactly 63 predictor variables after one-hot encoding, including engineered spatial features, pipe attributes, and selected categorical variables (e.g., fluid type, material, and flowline action). The resulting attributes retained and engineered are seen in **Table 1**.

These preprocessing steps formed the foundation of our analysis by ensuring the dataset's accuracy, relevance, and consistency for modeling. The addition of line age and risk variables directly supported our goal of identifying factors associated with flowline risks. By pruning and standardizing the data, we simplified the analytical process and enhanced the robustness of our results. These refined datasets were then prepared for predictive modeling to effectively assess spill risks. It was important to recognize that despite careful preparation, our dataset may still have contained inherent errors or limitations. Acknowledging these imperfections highlighted the necessity of cautious interpretation and validation of predictive models. By transparently detailing our preprocessing challenges and decisions, we provided a framework for future studies working with similarly constrained datasets and emphasized the importance of interpreting results within these limitations.

Exploratory Data Analysis

In this research, we first performed an Exploratory Data Analysis (EDA) to uncover patterns, anomalies, and insights from spatial and frequency data related to flowline risk assessment across Colorado. EDA served as a fundamental phase in our study, allowing us to visually and statistically interpret the

Table 1: Attributes for Flowline Analysis

Attribute	Description	Units
Status	Operational status of the flowline	-
Flowline Action	Actions taken or required on the flowline	-
Location Type	Type of facility	-
Fluid Type	Type of fluid transported	-
Material	Construction material of the flowline	-
Diameter	Diameter of the flowline	Inches
Length	Length of the flowline	Feet
Maximum Operating Pressure	Max pressure flowline can withstand	PSI
Line Age	Age of the flowline	Years
Number of Lines	Number of line segments of Geo-Spatial location	-
Bounding Box Area	Boxed area of Geo-Spatial location	$Feet^2$
Root Cause Type*	Underlying cause of the spill	-

*Attribute included only for flowlines associated with spills.

distribution of risk, which was crucial for subsequent predictive modeling and decision-making processes (Komorowski et al., 2016). By mapping out risk levels and their distribution, we aimed to identify high-risk areas that required immediate attention and better understand the underlying factors contributing to these risks. This approach not only improved our understanding of the spatial dynamics but also facilitated a targeted response strategy based on empirical evidence.

Fig. 3 provided a visual representation of risk distribution across Colorado. Blue markers indicated areas classified as 'Low Risk,' while red markers denoted 'High Risk.' The concentration of blue markers across the state suggested that most regions were currently at low risk. However, there were notable clusters of high-risk areas, particularly around major urban centers. This spatial distribution was essential for regional planning and resource allocation, highlighting areas that might have required more intensive monitoring and preventive measures. For additional clarity, **Fig. 4** provided an enlarged view of the high-risk cluster in the northern Front Range, specifically highlighting areas around Boulder, Fort Collins, and Denver.

In addition to spatial risk mapping, we analyzed the correlation between variables to better understand feature interdependencies (see **Appendix A-1**). The correlation heatmap revealed a moderate positive relationship between features such as length, bounding box area, and number of lines, which was expected given their shared geometric basis. Operational and categorical features, including material, fluid type, and status, showed weak to no correlation with spatial features, supporting their independent contribution to the predictive model. This analysis guided our dimensionality reduction choices and validated the inclusion of diverse feature types in the model.

Risk Assessment by Operational Parameters

To assess how operational factors influenced spill risk, we analyzed the risk frequency distributions of fluid type, pipe material, diameter, operator number, line age, and maximum operating pressure, as shown in **Appendix B-1**.

Overall, approximately 99.04% of flowlines were classified as low risk, with only 0.96% labeled as high risk, underscoring the rarity of spills but the importance of identifying contributing conditions. In terms of fluid type, crude oil and crude oil emulsion flowlines showed a slightly elevated frequency of high-risk classifications compared to other types such as natural gas or produced water. Further investigation was conducted to identify the contributing factors to spill risk associated with different types of fluids. When examining pipe material, most categories showed very few high-risk classifications. However, PVC and pipes labeled as "Other" had slightly more high-risk cases than common materials like carbon steel, HDPE, or fiberglass. This may have indicated potential quality or durability differences, though the small sample size of non-standard materials limited definitive conclusions.

For diameter, the majority of high-risk classifications occurred in lines with larger diameters, particularly 8-inch and 12-inch pipes. This may have reflected the increased volume or pressure demands typically associated with these dimensions. Operator number analysis showed that most operators had exclusively low-risk flowlines. One exception was Operator 98220, which appeared to have a dispropor-

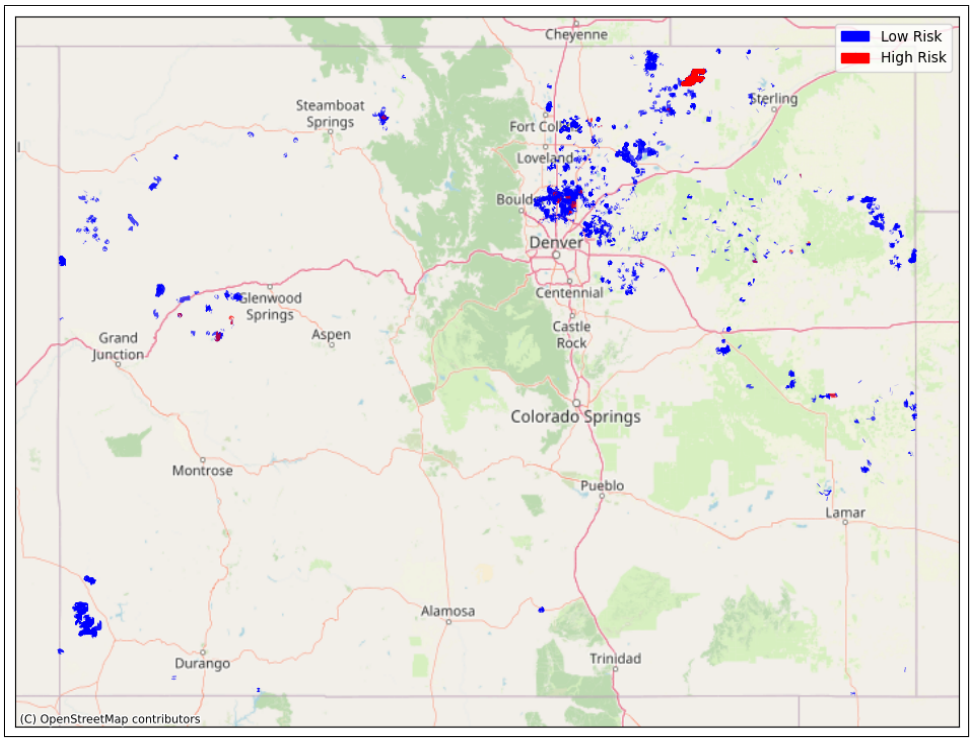

Fig. 3: Spatial Distribution of Flowline Risk in Colorado

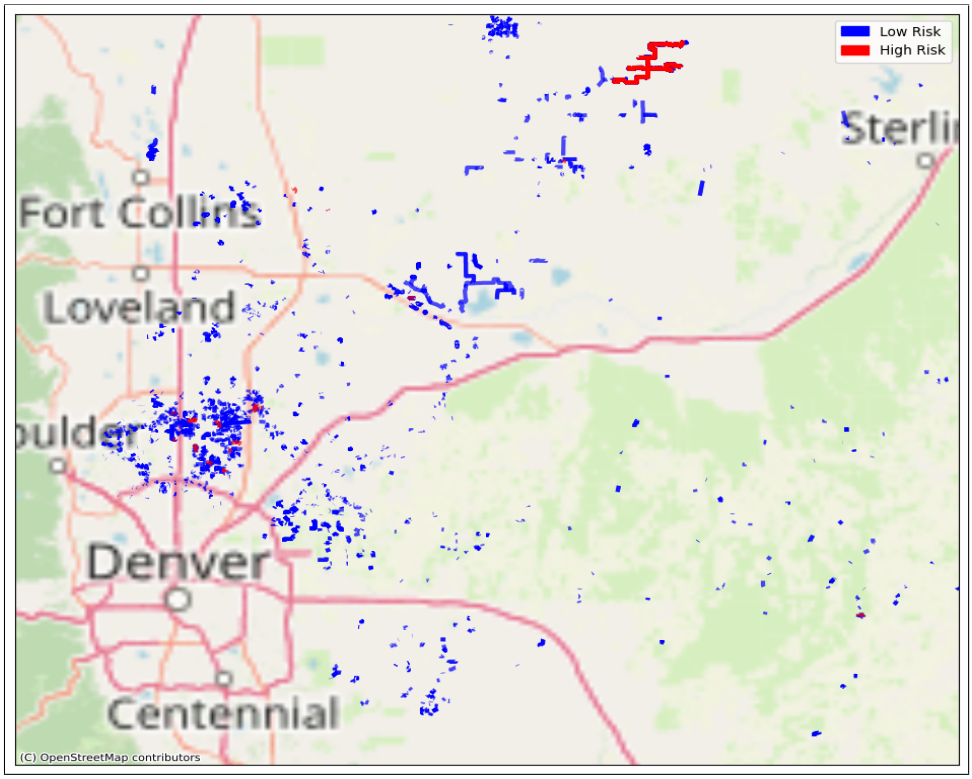

Fig. 4: Zoomed in Spatial Distribution of Flowline Risk in Colorado

tionately high-risk rate in our dataset. However, this was likely due to the small sample size—only two injection-related incidents were reported (in 2017 and 2019)—and potentially incomplete data matching. These conditions could have skewed apparent risk metrics and should not be interpreted as representative of actual operational risk. To quantify flowline ownership across operators, we calculated descriptive statistics based on the number of flowlines per operator. Flowline counts ranged from 1 to 837, with a median of 6 and a mean of 47.25. While there were 88 operators in total, the top 10 accounted for approximately 72% of all flowlines. This highlighted a concentration of flowline infrastructure among a small number of operators, which had implications for both model training and regulatory oversight.

In terms of line age, high-risk classifications increased marginally with older infrastructure. Most high-risk lines were over 50 years old, suggesting that age-related degradation may have played a role in spill risk. Lastly, maximum operating pressure presented a more nuanced pattern. While most pressure categories were dominated by low-risk flowlines, the frequency of high-risk cases was slightly higher in the 25 to 75 psi range and sporadically present up to 510 psi. This suggested that mid-range pressures may have been a useful indicator when evaluating risk, though the relationship was not strictly linear. These trends validated the inclusion of these variables in the predictive models and offered preliminary insight into potential indicators of elevated spill risk.

Methodology

In our study, we performed risk analysis using supervised ML techniques with an added application of Principal Component Analysis (PCA) (Hastie et al., 2009) and Canonical Correlation Analysis (CCA) (Hotelling, 1936) for dimensionality reduction. This methodology addressed the complexity of predicting spill risk in flowlines by applying various ML models to explore different aspects of the dataset. To prepare our dataset for ML applications, we first isolated the predictors (inputs) and the outcome variable (outputs). The inputs included all 64 features of interest from our dataset except for risk. The risk column was designated as our target variable, indicating the presence or absence of a spill.

The division of the dataset into training and testing sets was done using a stratified split, a technique that ensured each subset maintained the same ratio of spill to non-spill flowlines as in the original dataset. In other words, stratification guaranteed that the rare events—in this case, flowline spills, which accounted for only about 1% of the data—were proportionally represented in both the training and testing sets. This prevented situations where the minority class was over- or under-represented in one subset, which could otherwise have led to biased training or unreliable evaluation results. We allocated 70% of the total dataset for training and 30% for testing, resulting in approximately 2,910 flowlines in the training set and 1,248 in the test set.

Following the train-test split, we standardized the input features by adjusting each feature to have a mean of zero and a standard deviation of one. Standardization was a crucial preprocessing step, especially for models that were sensitive to feature scaling and techniques involving dimensionality reduction. We fit the scaler on the training data to learn the scaling parameters and then applied the same transformation to the test data to maintain consistency and prevent data leakage. This step ensured model stability, interpretability, and overall predictive performance.

Although cross-validation was not applied in this study due to computational constraints and the challenge of maintaining class balance across folds, we used a fixed random seed to ensure reproducibility. Predictor variables used in the model were limited to engineered features and encoded categorical attributes; again, non-informative identifiers such as operator number, flowline ID, and location ID were excluded to prevent overfitting. One-hot encoding increased the number of predictors from 12 to 64. All model performance metrics were computed on the held-out test set, and comparisons with training performance indicated consistent results across most classifiers, suggesting minimal overfitting.

Machine Learning Models

In our analysis of flowline spill risks, we employed a diverse array of supervised ML models. Within the supervised category, we further divided the techniques into single classifiers and ensemble classifiers, each with specific strengths and approaches to handling the prediction tasks. For the single classifiers, we chose three methods. Logistic Regression (LR) is a fundamental statistical approach that modeled the probability of a binary outcome based on one or more predictor variables (Hastie et al., 2009). It is particularly useful for understanding the impact of several independent variables on a binary response, making it ideal for binary classification tasks such as predicting whether a spill would occur or not.

Table 2: Top 5 Features Contributing to Principal Components 1 and 2

PC1		**PC2**	
Feature	Loading	Feature	Loading
Length (ft)	0.389606	Location Type - Production Facilities	0.308306
Bounding Box Area	0.383689	Fluid Type - Natural Gas	−0.277665
Number of Lines	0.375131	Number of Lines	0.275628
Diameter (in)	0.333562	Bounding Box Area	0.269872
Fluid Type - Crude Oil	0.318440	Material - Poly	−0.267661

K-Nearest Neighbors (K-NN) was a non-parametric method used for classification by comparing feature similarity. A data point is classified by a majority vote of its neighbors, with the data point being assigned to the class most common among its k nearest neighbors (Hastie et al., 2009). K-NN is highly intuitive and straightforward but could become computationally expensive as the size of the data grew. Support Vector Machine (SVM) is a powerful classification technique that found the hyperplane that best divided a dataset into classes. It is effective in high-dimensional spaces and robust against overfitting, especially in complex domains where the margin between different classes is clear (Hastie et al., 2009).

For the ensemble classifiers, we also chose three methods. Gradient Boosting Decision Trees (GBDT) is an ensemble technique that builds the model in a stage-wise fashion. It constructs new models that predict the residuals or errors of prior models and then combines them into a final model (Hastie et al., 2009). It's often used for its high performance and predictive power, particularly in competitions and real-world applications. Adaptive Boosting (AdaBoost) is one of the first boosting algorithms to be adapted to solving practices. It works by combining multiple weak classifiers to create a strong classifier. During the training phase, AdaBoost assigns weights to each instance, which are adjusted as each successive model is built, focusing more on difficult to classify instances (Hastie et al., 2009). Random Forests (RF) construct a multitude of decision trees at training time and output the class that is the mode of the classes (classification) of the individual trees. It is highly versatile and capable of handling both regression and classification tasks, offering a good balance between accuracy and computational efficiency (Breiman, 2001).

Dimensionality Reduction Analyses

While classification was the primary goal, PCA was also used as an exploratory tool to evaluate feature redundancy and support clustering analysis. PCA reduces the dimensionality of a dataset by identifying directions—known as principal components—that capture the maximum variance in the input features. Standardizing the dataset helped balance contributions from all variables and enabled us to identify more meaningful principal components.

The first two principal components explained only 13.1% of the total variance—7.5% by PC1 and 5.6% by PC2—which was expected in datasets with many one-hot encoded variables, where variance was spread thinly across numerous binary features. A cumulative explained variance plot (see **Appendix C-1**) showed that more than 40 components were required to explain 90% of the variance, which was neither practical nor interpretable for modeling. This reinforced our decision to focus on only the top two components for dimensionality reduction and visualization.

As shown in **Table 2**, PC1 was largely influenced by geometric and structural attributes such as flow-line length, bounding box area, number of lines, and diameter, as well as the presence of crude oil as the transported fluid. This suggested that PC1 captured variation related to spatial scale and infrastructure complexity. In contrast, PC2 reflected a mix of categorical and physical attributes, including facility type (production facilities), transported fluid type (natural gas), material type (poly), and again, number of lines and bounding box area. This component appeared to capture variation in operational context and construction materials. Despite the relatively low variance captured, these principal components were useful in revealing underlying data structure and informed subsequent unsupervised learning tasks.

In addition to PCA, we explored CCA as a supervised dimensionality reduction technique. CCA reduces dimensionality by identifying linear combinations of features in two datasets (X and Y) that are maximally correlated with each other. Unlike PCA, which focuses only on variance within the input features, CCA leverages information from both the predictor variables and the response variable to find

Table 3: Top 5 Features Contributing to Canonical Component 1

CC1	
Feature	Loading
Location Type - Produced Water Transfer System	0.473684
Flowline Action - Unknown	0.428694
Root Cause Type - Unknown	0.415709
Root Cause Type - Incorrect Operation	0.341643
Root Cause Type - Natural Force Damage	0.332114

directions that are directly aligned with the modeling objective. This distinction made CCA particularly valuable when the goal is to align the data transformation with a specific predictive target.

Although the risk column was binary, we treated it as a soft label, acknowledging that while spills labeled as 1 represented confirmed events, those labeled 0 may still have included unreported or undocumented spills. This approach allowed us to model risk as a continuous variable that reflected varying levels of spill risk, rather than a strict binary classification. Because risk was a single-output variable, we computed only one canonical component. CCA required both the predictor and response matrices to be two-dimensional, so we reshaped the target to meet this requirement while retaining its univariate structure. The result was a one-dimensional supervised projection of the input features that was maximally correlated with spill likelihood. This projection was then used as a continuous input in downstream classification models.

Table 3 presented the top five features contributing to the first canonical component (Canonical 1), which was derived through Canonical Correlation Analysis (CCA). This component represented a supervised linear projection of the input features that was most strongly correlated with the risk variable. The values listed reflected each feature's contribution, or loading, to this projection—higher values indicated stronger influence on the component and, by extension, greater alignment with spill likelihood.

The results showed that the first canonical component was most strongly influenced by categorical features related to infrastructure type and data completeness. The top contributing feature was location type (produced water transfer system), suggesting that flowlines serving these systems were more likely to be associated with spill events. Other high-loading features included Unknown values for both flowline action and root cause type, indicating that poorly documented or ambiguous operational details may have correlated with elevated risk—potentially due to oversight, lack of monitoring, or incomplete reporting. Additional contributors such as Incorrect Operation and Natural Force Damage as root cause types further emphasized the importance of operational failures and environmental stressors. Together, these results suggested that both physical infrastructure characteristics and metadata uncertainty were key indicators in identifying flowlines that may have been more prone to spills.

Each of these methods brought different strengths and weaknesses to the table. By integrating multiple approaches—from simple models like logistic regression to more complex ensemble methods—and leveraging the exploratory power of dimensionality reduction analyses, we aimed to capture a broad spectrum of insights from the dataset, providing a robust analytical framework for effectively assessing and predicting spill risks.

Results

In ML, particularly in classification tasks, several key performance metrics were crucial for evaluating the effectiveness of models. These metrics included accuracy, precision, recall, and the F1 score, each offering unique insights into a model's performance.

Accuracy measures the proportion of total correct predictions (both true positives and true negatives) among all cases evaluated. It provides a straightforward indicator of a model's overall effectiveness across all classes. However, its reliability can diminish in the presence of unbalanced datasets where one class significantly outnumbers another. Precision assesses the accuracy of positive predictions, that is, the proportion of true positive results among all positive predictions made by the model. High precision is vital in scenarios where the consequences of false positives are significant, ensuring that when a model predicts a risk, such a prediction is likely to be correct. Recall, or sensitivity, measures the model's ability to identify all actual positives from the data. It represents the proportion of true positives identified

Table 4: Without Dimensionality Reduction: Classifier Performance Metrics

Classifier	Accuracy	Precision	Recall	F1 Score
Single Classifiers				
LR	0.9976	0.91	0.83	0.87
K-NN	0.9976	1.00	0.75	0.86
SVM	0.9928	1.00	0.25	0.40
Ensemble Classifiers				
GBDT	0.9976	1.00	0.75	0.86
AdaBoost	0.9944	1.00	0.42	0.59
RF	0.9944	1.00	0.42	0.59

relative to the total actual positives. This metric is critical in situations where missing a positive instance carries serious consequences, such as failing to identify a potential hazard in risk management systems. The F1 Score harmonizes precision and recall by calculating their harmonic mean. This metric is particularly useful when seeking a balance between precision and recall, especially in environments with uneven class distributions. It ensures that a model does not overly favor one metric over the other, providing a more balanced view of model performance.

In the context of this study on flowline risk in the oil and gas industry, these metrics collectively informed the robustness of various ML models used to predict spill risks. High scores in these metrics indicated a model's capability to accurately predict and manage potential risks, which was crucial for ensuring safety and operational efficiency. The results from the ML models used in our study to predict spill risks in flowlines revealed significant insights into the effectiveness of each classifier under different circumstances, such as before and after applying PCA and CCA dimensionality reduction methods.

Without Dimensionality Reduction

Table 4 showed the results of our supervised ML models trained on the full feature set without PCA. LR and GBDT exhibited high accuracy and strong F1 scores, effectively identifying spill occurrences while minimizing false alarms. K-Nearest Neighbors (K-NN) achieved excellent precision but with slightly reduced recall, indicating reliable but cautious predictions of spills. SVM demonstrated high precision but notably lower recall and F1 score, suggesting challenges in consistently detecting the minority class instances. Among ensemble classifiers, GBDT performed robustly similar to LR, whereas RF and AdaBoost showed high precision yet lower recall, highlighting their limited ability to fully capture all spill cases. These results indicated that models trained on the complete set of engineered and categorical features remained effective for spill prediction tasks, particularly emphasizing the balanced performance of LR and GBDT in accurately detecting rare events.

PCA Dimensionality Reduction

Table 5 showed the results of our supervised machine learning models after applying PCA for dimensionality reduction. In contrast to models trained on the full feature set, single classifiers—LR, K-NN, and SVM—maintained high overall accuracy but demonstrated poor performance in detecting the minority class (spills). LR, notably, failed entirely to identify positive cases, resulting in zero precision, recall, and F1 scores. K-NN and SVM exhibited limited predictive capability, achieving precision values of 0.75 and 1.00, respectively, yet both suffered from very low recall (0.25), reflected in their low F1 scores (0.38 and 0.40). Ensemble classifiers similarly experienced a degradation in performance following PCA reduction. GBDT produced a modest F1 score of 0.30, indicating limited ability to classify spill cases effectively. RF matched the accuracy of K-NN (0.9920) and similarly achieved a precision of 0.75, recall of 0.25, and an F1 score of 0.38. AdaBoost demonstrated relatively high accuracy (0.9912) but low recall (0.17), resulting in a modest F1 score of 0.27. These results highlighted PCA's limitation in retaining crucial discriminative information, particularly in highly imbalanced datasets, reinforcing the necessity for careful feature selection and model interpretability to effectively capture minority class signals.

Table 5: With PCA Reduction: Classifier Performance Metrics

Classifier	Accuracy	Precision	Recall	F1 Score
Single Classifiers				
LR	0.9896	0.00	0.00	0.00
K-NN	0.9920	0.75	0.25	0.38
SVM	0.9928	1.00	0.25	0.40
Ensemble Classifiers				
GBDT	0.9888	0.38	0.25	0.30
AdaBoost	0.9912	0.67	0.17	0.27
RF	0.9920	0.75	0.25	0.38

Table 6: With CCA Reduction: Classifier Performance Metrics

Classifier	Accuracy	Precision	Recall	F1 Score
Single Classifiers				
LR	0.9976	0.91	0.83	0.87
K-NN	0.9976	0.91	0.83	0.87
SVM	0.9976	0.91	0.83	0.87
Ensemble Classifiers				
GBDT	0.9968	0.83	0.83	0.83
AdaBoost	0.9976	0.91	0.83	0.87
RF	0.9968	0.83	0.83	0.83

CCA Dimensionality Reduction

Table 6 presented the performance of our supervised machine learning models after applying CCA for dimensionality reduction. In contrast to the significant degradation observed with PCA, models trained on CCA-reduced features maintained strong predictive capabilities across all performance metrics. Notably, single classifiers—LR, K-NN, and SVM—achieved identical accuracy (0.9976) and robust precision, recall, and F1 scores (all at 0.91, 0.83, and 0.87, respectively). This consistency highlighted CCA's effectiveness in preserving class-discriminative information, especially for the minority class (spill events). Ensemble classifiers exhibited similarly strong performance with minimal drop in metrics. GBDT, RF, and AdaBoost maintained high accuracy (ranging from 0.9968 to 0.9976) while balancing precision and recall. AdaBoost mirrored the strong performance of single classifiers (F1 score of 0.87), while GBDT and RF achieved slightly lower but still impressive F1 scores of 0.83. These results suggested that CCA was particularly well-suited for preserving the predictive structure of the original feature set, even under dimensionality reduction. Compared to PCA, CCA enabled more effective learning by maintaining correlations between the original and reduced features, allowing both single and ensemble classifiers to detect rare spill events with high reliability.

Conclusion

This research demonstrated the substantial potential of integrating Machine Learning (ML) and Geographic Information Systems (GIS) for analyzing flowline risk within the oil and gas sector. Despite challenges related to data availability and tabularization of spatial information, our approach showed considerable promise in predicting spill risks. A notable contribution of this study was the development of a strategy for addressing sparse spill data by stratifying the train-test split, which was crucial given the frequent underreporting of spill events and the resulting class imbalance. We also explored dimensionality reduction techniques such as Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) to assess their effect on model performance in high-dimensional feature spaces, offering insight into the trade-offs between computational efficiency and predictive accuracy. Finally, we introduced a novel method for incorporating multilinestring GeoDataFrames into ML models by extracting

quantifiable spatial features—an approach less explored compared to traditional raster data applications. Together, these approaches helped mitigate class imbalance, manage high-dimensional data, and unlock the predictive value of complex spatial datasets in large-scale oil and gas infrastructure risk assessments.

The Exploratory Data Analysis (EDA) validated our feature selections, such as pipe age, diameter, fluid type, material, and root cause type. While these variables contributed valuable insights, root cause data was sparse in earlier records, which restricted the depth of failure-mode analysis in this study. However, it was worth noting that data quality in this area had been steadily improving. In particular, ECMC's Engineering Integrity team had strengthened root cause documentation through more stringent reviews of Form 19 supplemental reports and had increasingly followed up with operators to verify reported causes. As this process matured, future analyses would likely benefit from higher-quality root cause attributions. Nevertheless, the EDA in this study effectively highlighted regions and variables significantly correlated with risk, providing a strong foundation for predictive modeling. Supervised ML models exhibited robust performance using the complete feature set, with Logistic Regression (LR) and Gradient Boosting Decision Trees (GBDT) standing out with balanced precision, recall, and high F1 scores. These classifiers effectively identified spill occurrences while maintaining low false-positive rates.

Conversely, applying PCA led to diminished predictive capabilities, especially for LR, which completely failed to detect spills under dimensionality reduction. K-Nearest Neighbors (K-NN) and Support Vector Machines (SVM) also faced significant performance degradation after PCA, underscoring the challenges of maintaining discriminative information in highly imbalanced datasets. PCA visualizations further validated these clusters, though they indicated nuances in the clustering process relative to ground truth labels, suggesting opportunities for refinement in future analyses.

In contrast, CCA preserved model performance remarkably well following dimensionality reduction. All single and ensemble classifiers maintained high accuracy, precision, recall, and F1 scores—closely matching or even equaling those achieved without any reduction. For example, LR, K-NN, and SVM each achieved an F1 score of 0.87 with CCA, identical to their performance on the full feature set. Ensemble methods like GBDT and Random Forest also demonstrated balanced recall and precision, indicating that the reduced feature space retained sufficient discriminative power for detecting spill events. This suggested that, unlike PCA, CCA was capable of maintaining the underlying structure necessary to effectively classify rare events in imbalanced datasets. By aligning linear combinations of feature sets in a way that maximized correlation, CCA proved to be a compelling alternative when dimensionality reduction was needed without sacrificing model interpretability or minority class performance.

In comparing the three modeling approaches—without dimensionality reduction, with PCA, and with CCA—we found that CCA offered a strong middle ground. While PCA significantly hindered model performance due to its inability to retain key class-distinguishing information, CCA preserved the predictive power of the original feature set even after reduction. In fact, models using CCA achieved performance nearly identical to those trained without dimensionality reduction. Both single classifiers and ensemble classifiers maintained high F1 scores along with balanced precision and recall. This suggested that CCA was a promising dimensionality reduction technique for high-dimensional, imbalanced classification problems, providing computational benefits without compromising classification accuracy.

Despite the results, our study acknowledged several limitations, particularly regarding data completeness and quality. Important variables like soil properties and operational logs were unavailable, restricting the attribution of spill risks to specific physical or environmental drivers. Moreover, spatial uncertainties required adopting a conservative 25-meter matching threshold, which significantly reduced usable spill records and potentially introduced sampling biases. However, we recognized that more accurate spill-to-flowline associations might have been achievable with targeted review of Form 19 supplemental reports and additional operator communications. In future work, we planned to collaborate directly with ECMC subject matter experts to refine this matching process and improve attribution confidence. Future research should also prioritize acquiring comprehensive datasets that include detailed physical characteristics, spatially precise spill records, and temporal operational data. Incorporating digital elevation models (Balasubramanian, 2017) and advanced segmentation techniques would enable more precise spatial predictions, and integrating temporal data could improve forecasting capabilities and proactive risk management.

By advancing predictive modeling and spatial analysis techniques, this research set a foundation for enhanced flowline risk assessment, enabling industry stakeholders to anticipate failures more accurately and respond proactively. The integration of machine learning with geographic information systems not only supported safer and more sustainable oil and gas operations but also provided a replicable framework adaptable to other infrastructure risk management challenges. Moving forward, continued collaboration and richer data acquisition would further strengthen these methods, ultimately safeguarding

environments, protecting communities, and driving innovation in industrial safety practices.

Finally, the code to reproduce our findings was made publicly available at `https://github.com/ichittumuri/SPE_Journal_Risk_Analysis_of_Flowlines`, and the datasets used could be accessed via the ECMC public data portal: `https://ecmc.colorado.gov/data-maps/downloadable-data-documents`.

Nomenclature

- **AdaBoost** = Adaptive Boosting – an ensemble learning method for improved accuracy
- **CCA** = Canonical Correlation Analysis – supervised projection method
- **ECMC** = Colorado Energy and Carbon Management Commission – the data source in this study
- **ESPG:26913** = A specific Coordinate Reference System (CRS) for mapping
- **GBDT** = Gradient Boosting Decision Trees – an ensemble classifier used in risk analysis
- **GIS** = Geographic Information Systems – used to analyze and visualize spatial data
- **K-NN** = K-Nearest Neighbors – a non-parametric classification method
- **LR** = Logistic Regression – a binary classification model used in this study
- **ML** = Machine Learning – used to predict risk in flowlines
- **PCA** = Principal Component Analysis – unsupervised variance-based reduction
- **RF** = Random Forest – an ensemble method used in classification tasks
- **SVM** = Support Vector Machine – a classification algorithm used for flowline risk analysis

List of Figures

Acknowledgments

This work was supported by the Mark Martinez and Joey Irwin Memorial Public Projects Fund: Design and Statistical Analysis of a Flowline Risk Assessment Model. The authors acknowledged the Colorado Energy and Carbon Management Commission (ECMC) for providing the data essential to this study. We also thanked Mark Schlagenhauf, P.E., Engineering Integrity Supervisor at ECMC, for his guidance and feedback throughout the project. Lastly, special thanks went to Matthew Bauer, Affiliate Faculty at the Colorado School of Mines, for his valuable GIS contributions and support during the research process.

Appendix A: Correlation Heatmap of Flowline Features

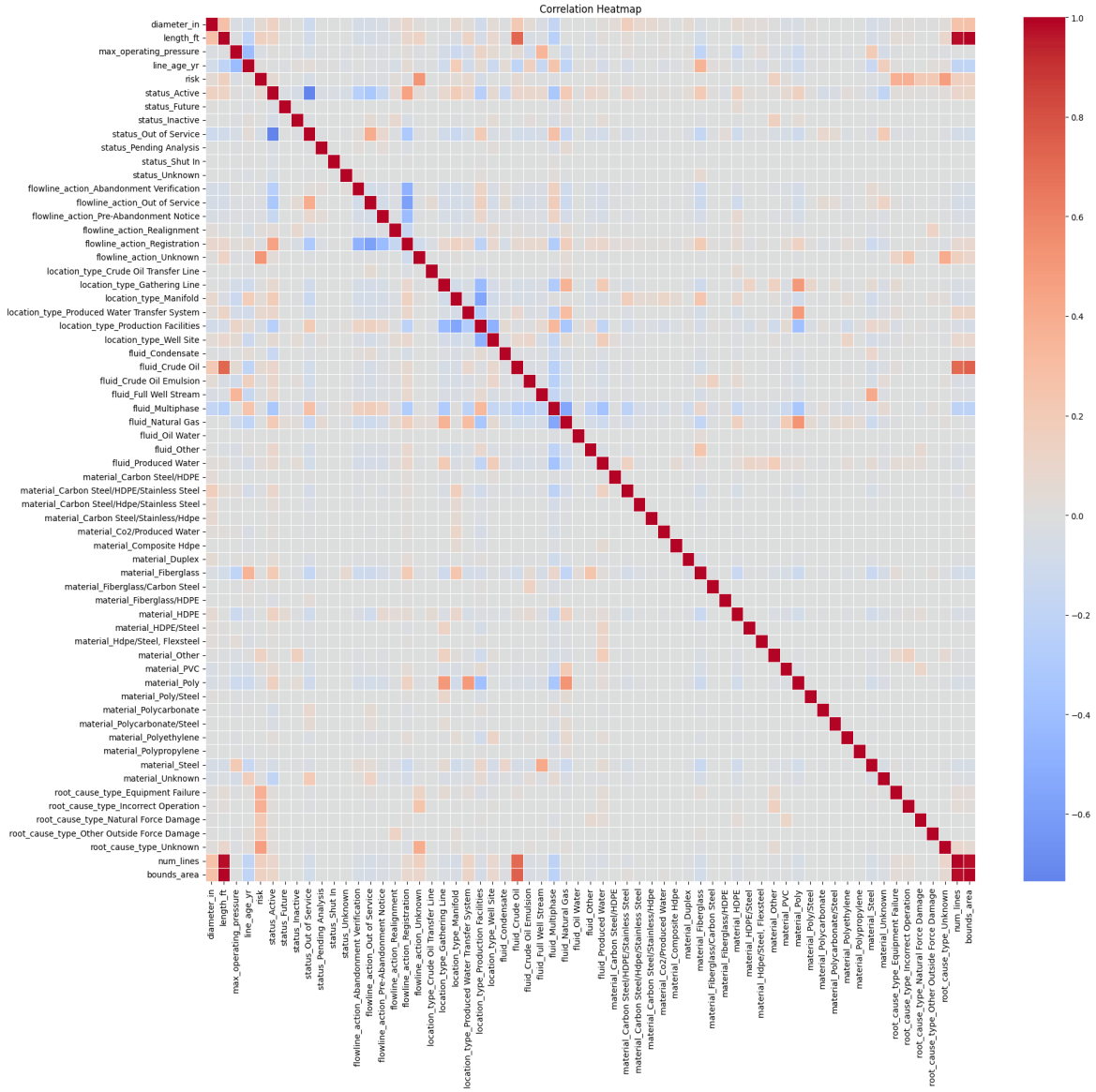

Fig. A-1: Correlation heatmap showing Pearson coefficients between flowline features.

Appendix B: Risk Breakdown Figures

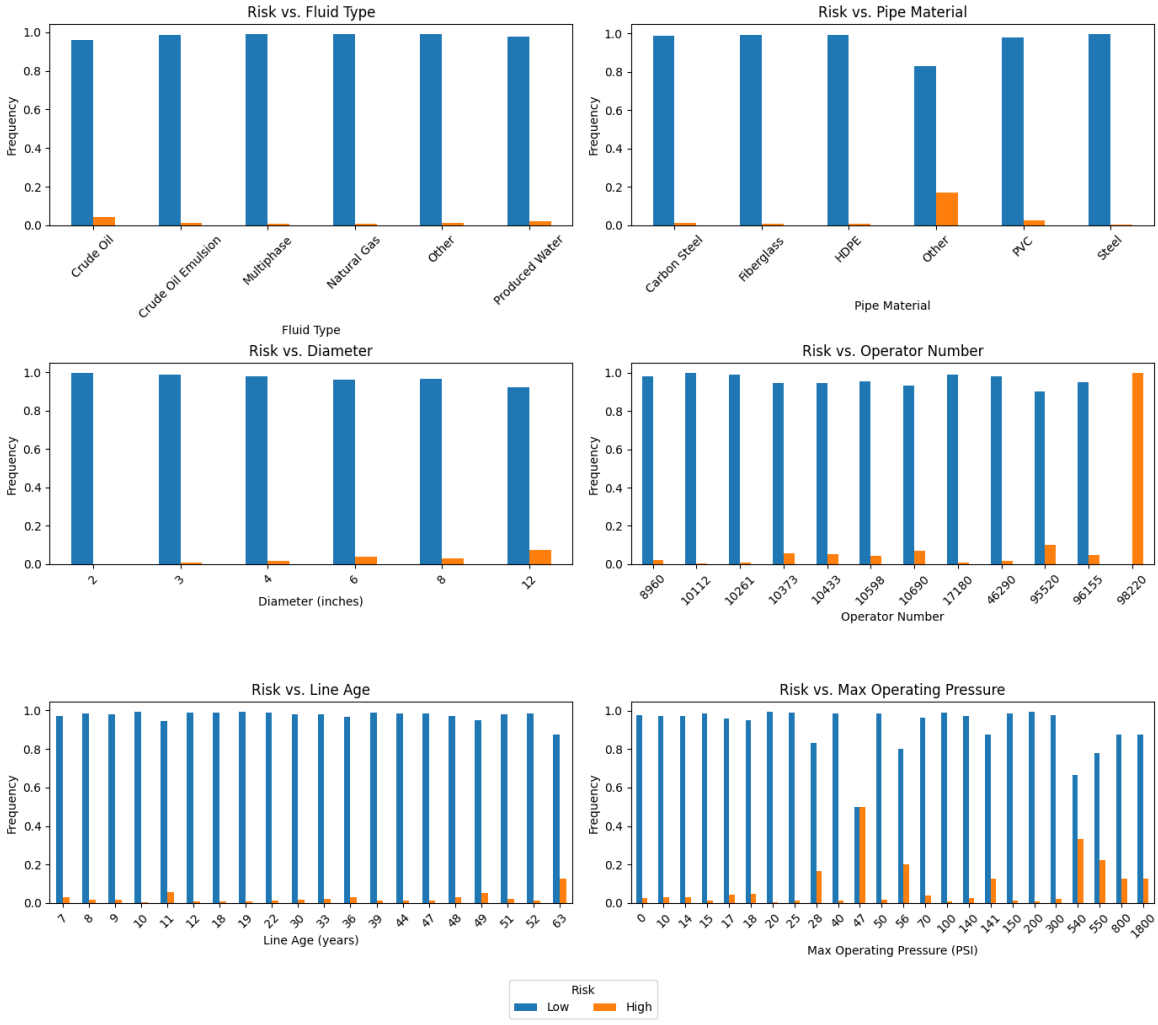

Fig. B-1: Visual breakdown of risk across operational parameters.

Appendix C: PCA Visualization

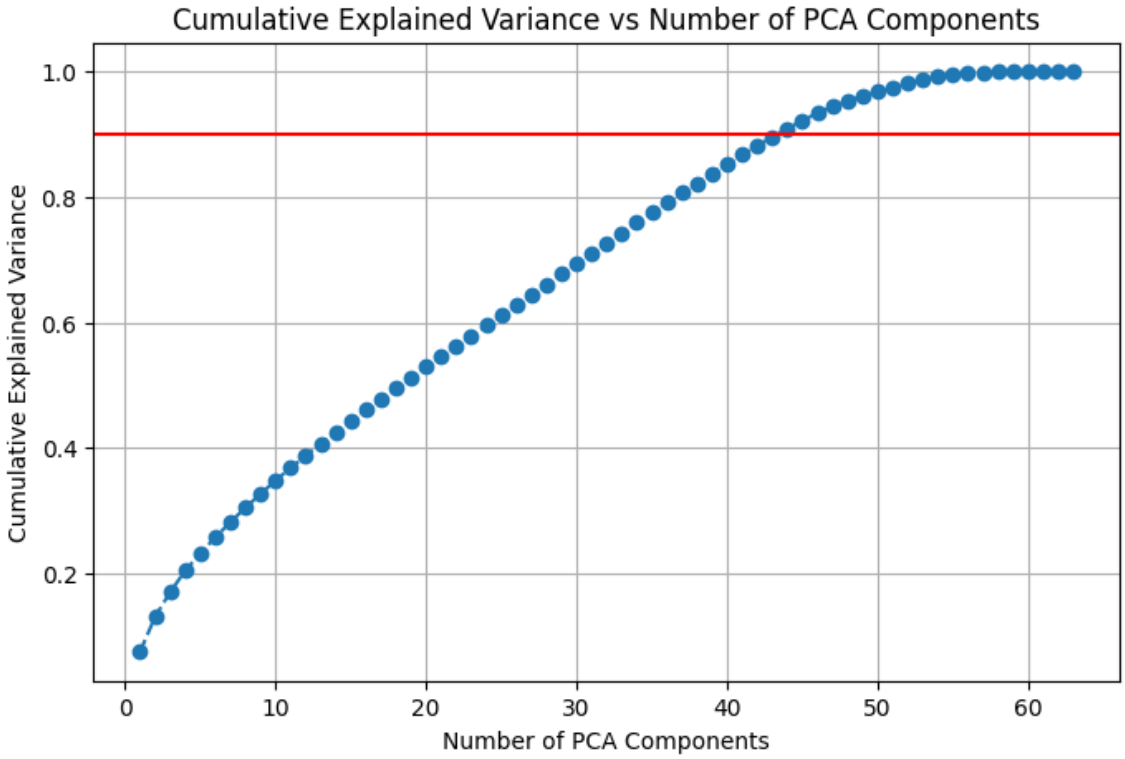

Fig. C-1: Principal Component Analysis (PCA) component visualization.

References

Balasubramanian, A. (2017). *Digital Elevation Model (DEM) in GIS*. University of Mysore.

Breiman, L. (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Fleckenstein, W. (2018). *Flowline Risk Review – Final Report*. Tech. rep. Presented to Mark Schlagenhauf and Stuart Ellsworth, October 2018. Colorado School of Mines.

Guan, S., F. Ayello, N. Sridhar et al. (2019). "Application of Probabilistic Model in Pipeline Direct Assessment". In: *CORROSION 2019*. Nashville, TN, USA.

Hastie, T., R. Tibshirani J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY, USA: Springer. DOI: 10.1007/978-0-387-84858-7.

Hotelling, H. (1936). "Relations between Two Sets of Variates". In: *Biometrika* 28.3/4, pp. 321–377.

Khalilpasha, H., J. Brown et al. (2023). "Minimising Pipeline Leaks and Maximising Operational Life by Application of Machine Learning at Cooper Basin". In: *AMPP Annual Conference + Expo*. Denver, CO, USA.

Komorowski, M., D. C. Marshall, J. D. Salciccioli et al. (2016). "Exploratory Data Analysis". In: *Secondary Analysis of Electronic Health Records*. Cham: Springer. DOI: 10.1007/978-3-319-43742-2_15.

Lee, L. H., R. Rajkumar, L. H. Lo et al. (2013). "Oil and Gas Pipeline Failure Prediction System Using Long Range Ultrasonic Transducers and Euclidean-Support Vector Machines Classification Approach". In: *Expert Systems with Applications* 40.6, pp. 1925–1934. DOI: 10.1016/j.eswa.2012.10.006.

Mazzella, J., T. Hayden, L. Krissa et al. (2019). "Estimating Corrosion Growth Rate for Underground Pipeline: A Machine Learning Based Approach". In: *CORROSION 2019*. Nashville, TN, USA.

Oil and Gas Conservation Commission, Colorado (2019). *Annual Flowline Spill Report – 2019*. Tech. rep. Denver, CO, USA: Colorado Oil and Gas Conservation Commission.

Rachman, A., T. Zhang R. M. Ratnayake (2021). "Applications of Machine Learning in Pipeline Integrity Management: A State-of-the-Art Review". In: *International Journal of Pressure Vessels and Piping* 193. DOI: 10.1016/j.ijpvp.2021.104471.

Senouci, A., M. Elabbasy, E. Elwakil et al. (2014). "A Model for Predicting Failure of Oil Pipelines". In: *Structure and Infrastructure Engineering* 10.3, pp. 375–387. DOI: 10.1080/15732479.2012.756918.

Vinogradov, P. V., K. V. Litvinenko, R. I. Valiakhmetov et al. (2018). "Development of a Model for Ranking Field Pipelines Based on Risk Assessment in Exploitation". In: *OIJ* 8, pp. 84–86. DOI: 10.24887/0028-2448-2018-8-84-86.

Zhang, Z. G., G. Liu et al. (2023). "A Prediction of Corrosion-Related Leakage on Distribution Pipelines via Machine Learning Method". In: *AMPP Annual Conference + Expo*. Denver, CO, USA.