# Risk Analysis of Oil and Gas Flowlines: A GIS and ML Approach

Isabella Chittumuri[1], Ryan Voss[1], Logan Douglass[1], Steve Wheeler[2], Soutir Bandyopadhyay[1]

[1]Department of Applied Mathematics and Statistics, Colorado School of Mines
[2]Colorado Energy and Carbon Management Commission (ECMC)

## Introduction

- Flowlines are underground pipes transporting oil, gas, and water.
- Failures can cause environmental harm and human risk.
- **Objective:** Develop predictive models to identify high-risk flowlines in Colorado using real-world regulatory data.

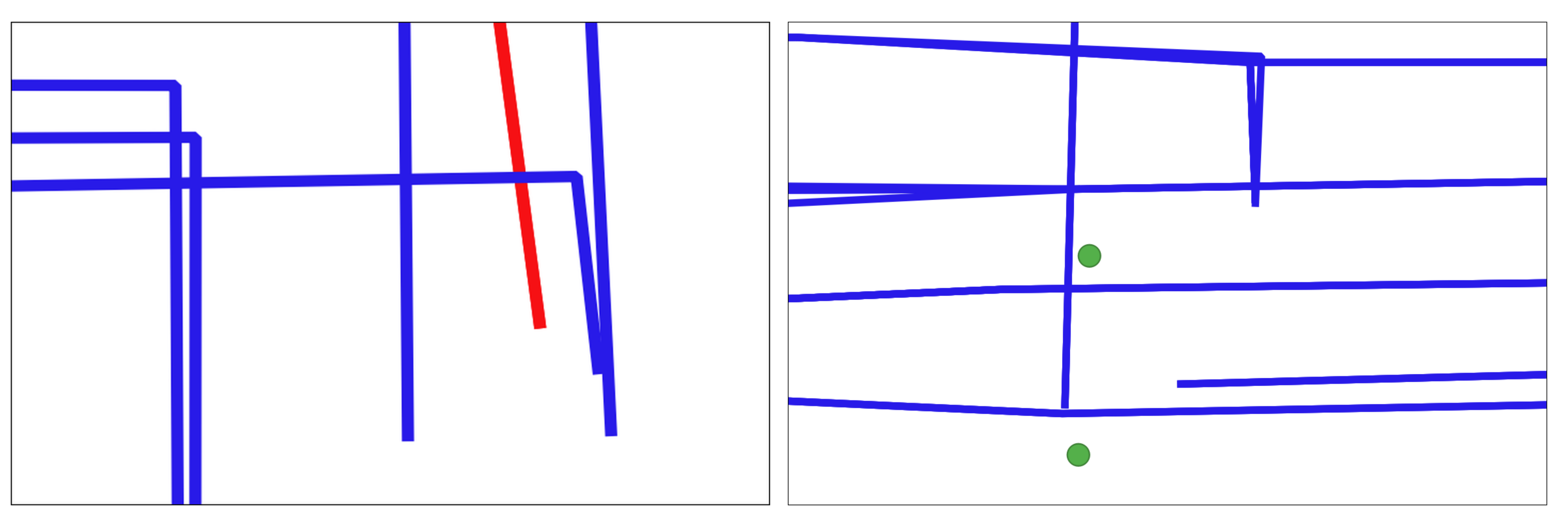
## Data Processing

Data Sources:

- 1,726 ECMC-reported spill events (2014–2022) [4]
- 21,000+ flowlines (spatial and operational attributes)

Spatial Matching:

- Operational and descriptive datasets merged by endpoints using a 25m tolerance.
- 4,117 flowlines were successfully matched.
- 41 spills were linked to specific flowlines with high confidence.



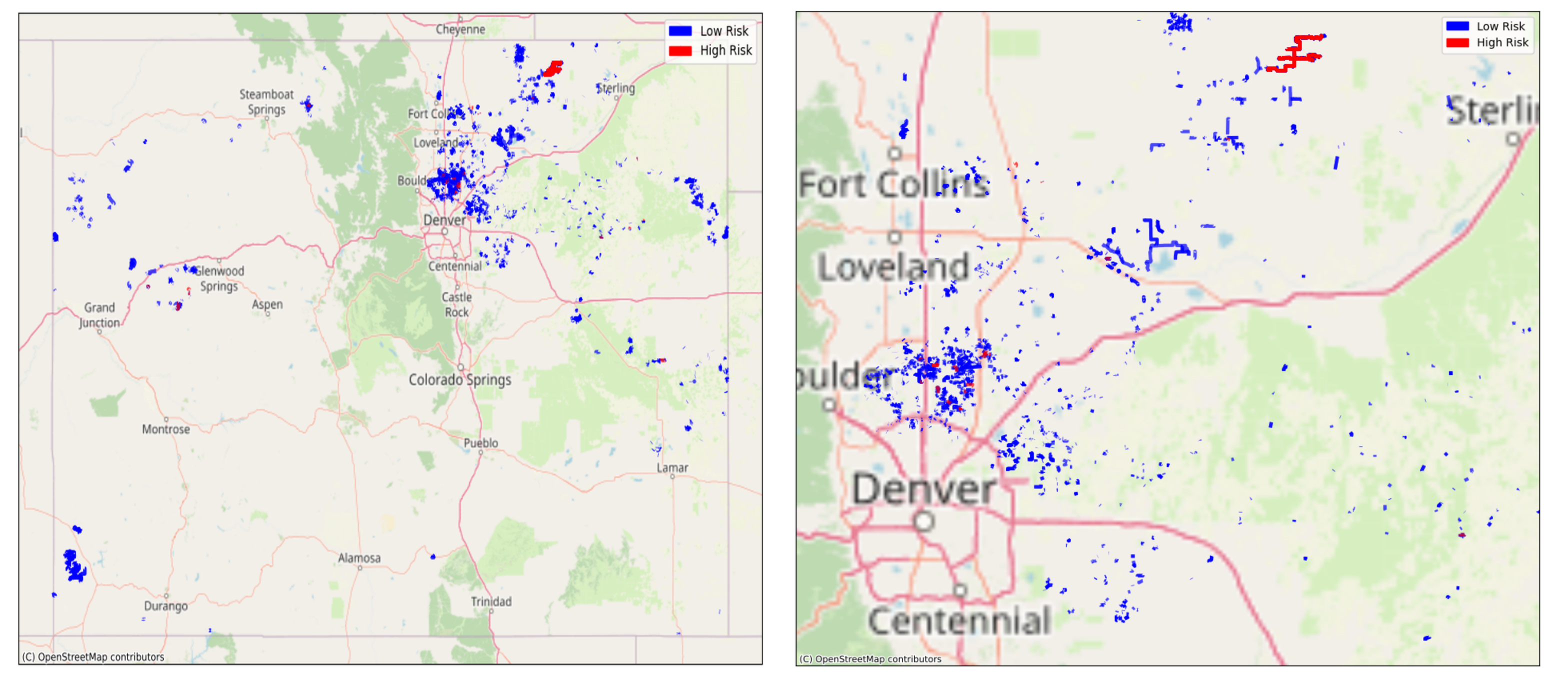
Feature Engineering:

- Created binary `risk` variable: 1 = spill, 0 = no spill.
- Spatial features extracted: line complexity, bounding box area, etc.
- Categorical variables one-hot encoded (e.g., material, fluid type).

## 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 |
| Max Operating Pressure | Max pressure flowline can withstand | PSI |
| Line Age | Age of the flowline | Years |
| Number of Lines | Number of line segments (geometry complexity) | - |
| Bounding Box Area | Area enclosing the flowline | $\text{Feet}^2$ |
| Root Cause Type* | Underlying cause of the spill | - |

## Observed Data




## Supervised ML Models

- Logistic Regression **(LR)**: Estimates spill probability using [2]:

$$P(Y=1 \mid \mathbf{X}) = \frac{1}{1+\exp(-z)}$$

- K-Nearest Neighbors **(KNN)**: Assigns class based on Euclidean distance:

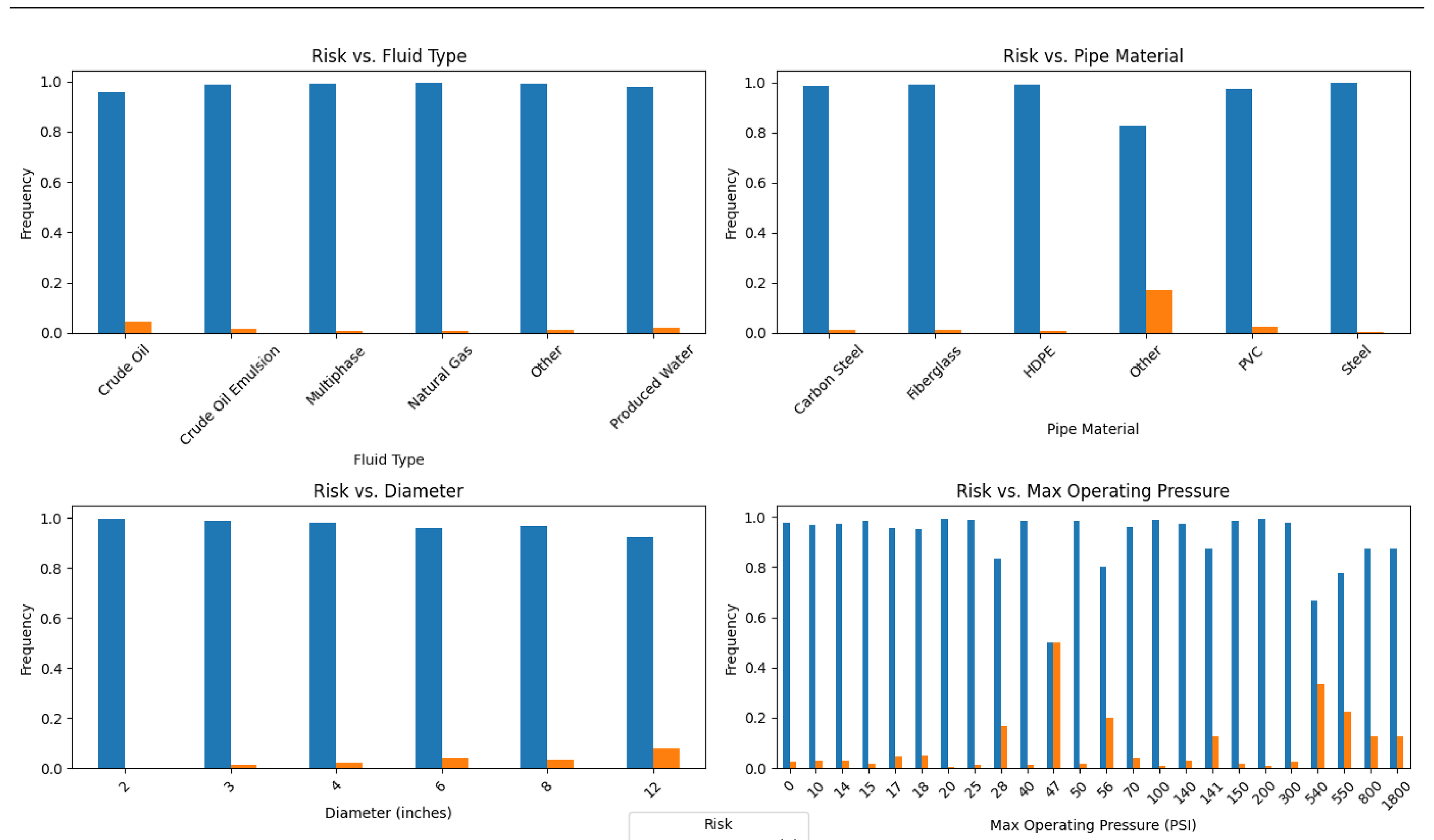
$$d(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{i,j})^2}$$

- Support Vector Machine **(SVM)**: Finds a hyperplane that maximizes the margin using [2]:

$$\hat{y} = \operatorname{sgn}(\mathbf{w}^\top \mathbf{x} + b)$$

- Gradient Boosting **(GBDT)**: Trains trees sequentially to reduce prior errors.
- Adaptive Boosting **(AdaBoost)**: Reweights data to emphasize hard-to-classify points.
- Random Forest **(RF)**: Averages predictions from trees fit on bootstrapped samples. [1]

## Risk Assessment of Operational Parameters




## Dimensionality Reduction Methods

- Principal Component Analysis (**PCA**): An *unsupervised* method that finds new axes (directions) capturing the most variance in the data. It solves:

$$\max_{\mathbf{w}} \ \operatorname{Var}(\mathbf{X}\mathbf{w}) \quad \text{s.t. } \|\mathbf{w}\| = 1$$

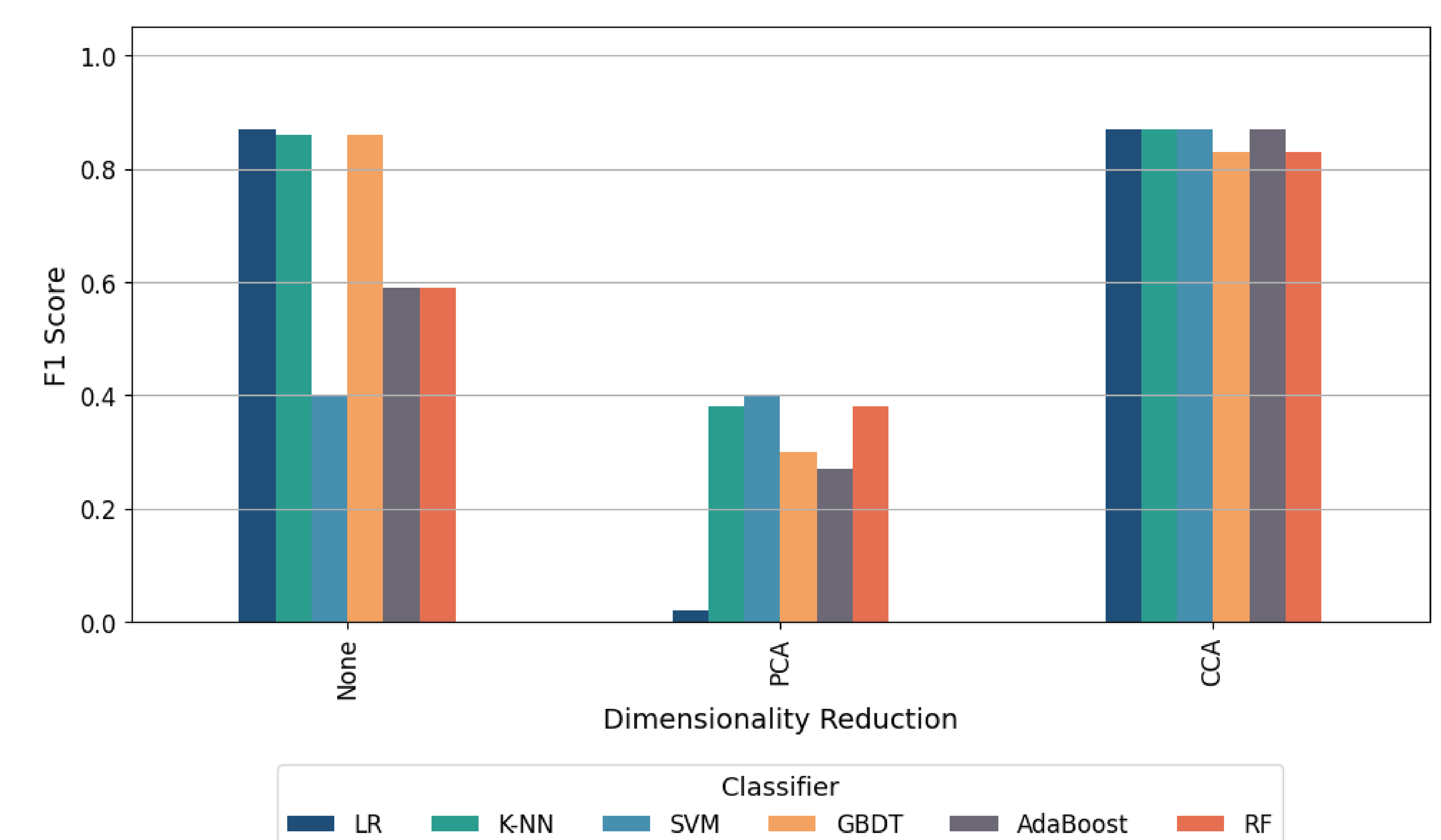
where $\mathbf{w}$ is a direction in feature space. Used to compress data while keeping the most information. [2]

- Canonical Correlation Analysis (**CCA**): A *supervised* method that finds combinations of features most correlated with the target. It solves:

$$\max_{\mathbf{a}} \ \operatorname{Corr}(\mathbf{X}\mathbf{a}, \mathbf{y})$$

where $\mathbf{X}$ is the feature matrix and $\mathbf{y}$ is the target. Great for focusing on patterns tied to prediction. [3]

## F1 Score Across Models and Reduction Methods




## Summary & Future Work

We benchmarked six classifiers across three reduction methods to predict spill risk. CCA preserved performance best; PCA reduced it.

- Incorporate spatial autocorrelation into model design using geostatistical methods.
- Expand to time-aware risk predictions and more granular flowline attributes.

## References

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, NY, USA, 2 edition, 2009.
[3] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.
[4] Oil and Gas Conservation Commission, Colorado. Annual flowline spill report – 2019. Technical report, Colorado Oil and Gas Conservation Commission, Denver, CO, USA, 2019.