ROSEN GROUP

Project 4: Predictive Maintenance with Survival Models

Complete Project Plan & Technical Documentation

Draft v1.0 — 3 February 2026

Prepared by:

Kent — Telecommunications Engineer / ML Practitioner

Heidenau, Saxony, Germany

1. Project Overview

1.1 Background

ROSEN Group is a global technology company headquartered in Stans, Switzerland, specializing in the development and operation of In-Line Inspection (ILI) tools for oil and gas pipelines. With operations in over 110 countries and 40+ locations worldwide, ROSEN is an industry leader in pipeline integrity management. Their ILI tools — commonly known as “smart pigs” — capture high-resolution data on pipe wall condition, including metal loss from corrosion, cracks, dents, and geometric anomalies.

The United States maintains approximately 300,000 miles of natural gas transmission pipelines, with 43% of this infrastructure installed during the 1950s–1960s. As this aging network approaches 60–75 years of service, operators face critical decisions about inspection scheduling, maintenance prioritization, and replacement planning. The Pipeline and Hazardous Materials Safety Administration (PHMSA) regulates this infrastructure and collects extensive data on both incidents (failures) and annual infrastructure reports from all operators.

1.2 Project Objective

Build a machine learning survival model that predicts the probability of pipeline failure (specifically corrosion-induced failure) as a function of pipe age, material properties, operational characteristics, and environmental conditions. The model will:

- Quantify the hazard rate (failure probability per unit time) for each pipe vintage and configuration
- Identify which pipe characteristics (coating era, diameter, %SMYS, class location) most strongly predict corrosion failure
- Generate operator-level risk scores that guide ILI inspection scheduling and prioritization
- Demonstrate the incremental value of combining ROSEN’s proprietary ILI data with public PHMSA records

1.3 Value Proposition for ROSEN

This project creates a data-driven framework that transforms ROSEN’s market positioning from reactive (“we inspect what you ask us to inspect”) to predictive (“our models identify which pipes need inspection most urgently”). The survival model provides three concrete business outcomes: (1) operators can optimize inspection intervals based on quantified risk rather than fixed schedules, reducing unnecessary inspections while catching high-risk segments earlier; (2) ROSEN can demonstrate, with statistical evidence, that ILI data improves failure prediction beyond what public records alone provide; and (3) the model serves as a technical differentiator in competitive proposals to pipeline operators.

2. Data Sources

2.1 PHMSA Flagged Incidents (Event Data)

Source: Pipeline and Hazardous Materials Safety Administration (PHMSA), U.S. Department of Transportation

URL: https://www.phmsa.dot.gov/data-and-statistics/pipeline/source-data

File: PHMSA_Pipeline_Safety_Flagged_Incidents.zip (25.3 MB)

Coverage: 21,134 incidents from 1986 to January 2026, across all pipeline types

This dataset contains every reportable pipeline incident filed with PHMSA. Each record includes the date, location, cause, operator, pipe characteristics (diameter, wall thickness, SMYS, coating type, seam type, material), and installation year. For this project we use the Gas Transmission & Gathering subset from 2010 to present: 1,985 incidents with 637 columns per record.

Key Subsets

Subset	Count	% of Total	Role in Model
All GT Incidents 2010–Present	1,985	100%	All-cause failure events
Corrosion Incidents	394	19.8%	Primary target variable
Material/Weld/Equipment	892	44.9%	Secondary target (ILI-detectable)
Excavation Damage	217	10.9%	External cause (not modeled)
Natural Force	146	7.4%	External cause (not modeled)

Data Quality for Corrosion Subset: 93% covariate coverage (diameter, wall thickness, SMYS, coating, seam type all present). Installation year available for 95% of corrosion incidents. Mean age at corrosion failure: 46.4 years (median: 46). This exceptional coverage makes corrosion the ideal modeling target.

2.2 PHMSA Annual Reports (Exposure / Denominator Data)

Source: PHMSA Annual Report Data, Gas Transmission & Gathering

URL: https://www.phmsa.dot.gov/data-and-statistics/pipeline/gas-distribution-gas-gathering-gas-transmission-hazardous-liquids

File: annual_gas_transmission_gathering_2010_present.zip (100.4 MB)

Coverage: 15 annual files (2010–2024), approximately 1,400 operators, 54 states/territories

Every pipeline operator in the United States is required by 49 CFR Parts 191 to submit annual reports to PHMSA detailing their infrastructure. These reports contain no incident information; instead, they provide the denominator for survival analysis — how many miles of pipe of each type are in service and exposed to risk. The reports are structured in multiple “Parts” (A through R), each capturing a different cross-section of the pipeline network.

Annual Report Parts Used in This Project

Part	Content	Survival Model Role	Key Variables
A–D	Operator info, commodity, total miles by	Covariable: material composition, operator	CP-coated, CP-bare, unprotected, plastic

	material (9 types)	size	
H	Miles by diameter (26 size brackets)	Covariable: diameter distribution	4" to 58"+, onshore/offshore
J	Miles by decade of installation (10 bins)	TIME VARIABLE + EXPOSURE: pipe age & miles at risk	Pre-1940 through 2020–29
K	%SMYS by class location (cross-tabulation)	Covariable: mechanical stress × population density	10 SMYS brackets × 4 class locations
M	Leaks/repairs by cause (13 categories)	Covariable: prior failure history	Ext corrosion, int corrosion, equipment, excavation

Schema Compatibility: Verified 100% identical column structure (117 columns) across all 15 years (2010–2024). This allows direct concatenation into a longitudinal panel without schema reconciliation.

2.3 Mendeley ILI Dataset (Supplementary)

Source: Mendeley Data — Probabilistic Degradation Models

File: probabilistic_degradation_models.zip (333 MB)

This academic dataset contains pipe-level In-Line Inspection data including individual metal loss measurements, wall thickness profiles, and degradation observations. It provides a bridge between the aggregate PHMSA data (operator-level) and the pipe-level detail that ROSEN's commercial ILI tools capture. It will be integrated in Phase 2 of the model as validation and for demonstrating the value of ILI-resolution data.

3. Work Completed (EDA Phase)

The Exploratory Data Analysis phase was conducted on 3 February 2026 across seven working sessions. All analyses were performed in Python using openpyxl, csv, and standard libraries. Results are documented in separate EDA reports and preserved in the survival panel CSV.

3.1 Session Timeline

Session	Focus	Key Output
1	ROSEN Group research & 4 ML project proposals	4 proposals with data sources identified
2	Data availability verification for all 4 projects	URLs, sizes, access methods confirmed
3	Project 4 selected; data download (14 files, 555 MB)	All datasets downloaded to C:\ PHMSA
4	Flagged Incidents EDA (21,134 incidents)	EDA report: cause distribution, age-at-failure, data quality
5	Annual Reports EDA (Parts A–M decoded)	EDA report: schema decode, miles by decade, denominators
6	Survival table construction (Part J + K + M joins)	Prototype table: 5,119 obs for 2024
7	Multi-year panel build + hazard rate analysis	7-year panel: 34,226 obs, 657 events, 106 corrosion

3.2 Flagged Incidents EDA

The complete incident dataset (21,134 records from 1986–present) was analyzed. We identified the Gas Transmission 2010–Present subset (1,985 incidents) as the primary event source. Key findings include: corrosion incidents have dramatically superior data quality (93% covariate coverage vs 46% for general incidents); the age distribution at failure shows a clear bimodal pattern with infant mortality at 0–9 years (18.4% of failures) and a degradation peak at 40–69 years (46.2%); and corrosion failures have a mean age of 46.4 years versus 38.6 for all causes combined, confirming that corrosion is a degradation-driven mechanism amenable to predictive modeling.

3.3 Annual Reports EDA

The 2024 Annual Reports were decoded across all Parts (A through M). Part J was identified as the critical denominator for survival analysis: it provides miles of pipe by decade of installation per operator per state per year, which when combined with incident data creates the exposure-at-risk measure essential for rate calculations. The national pipeline age profile was established: 43.2% of the U.S. gas transmission network was installed in the 1950s–1960s (130,000 miles now aged 55–74), representing the peak corrosion risk cohort. Part K provides a unique cross-tabulation of %SMYS by class location — the only joint distribution available in the public data — revealing that 110,000 miles operate at 61–72% SMYS in Class 1 (rural) locations. Part M provides leak/repair history by cause per operator, which serves as a powerful lagged predictor of future failures.

3.4 Survival Panel Construction

The survival analysis panel was constructed by joining Annual Reports (Part J: exposure) with Flagged Incidents (events) on the composite key (OPERATOR_ID, STATE_NAME, REPORT_YEAR, DECADE_BIN). Each observation represents one operator-state-decade combination in one year, with the count of incidents as the response variable and miles of pipe as the exposure offset.

Panel Dimensions (7-Year Build: 2017–2024)

Metric	Value
Total observations	34,226
Event observations (at least 1 incident)	657 (1.9%)
Censored observations (no incident)	33,569 (98.1%)
Total incidents matched	657
Corrosion incidents matched	106
Material/weld incidents matched	~276
Total pipe-miles covered	2,067,523
Unique operators	~1,400
States/territories	54
Report years available	7 (2017–2024, excluding 2019)

3.5 Key Empirical Finding: Non-Monotonic Corrosion Pattern

The central finding of the EDA phase is that corrosion failure rates are NOT monotonically related to pipe age. When hazard rates are computed by vintage (decade of installation), a striking non-monotonic pattern emerges:

Vintage	Miles (×4yr)	Corrosion Events	Rate /1K mi/yr	Interpretation
1970–79	202,328	15	0.074	HIGHEST — Coal tar coatings, early CP
Pre-1940	57,280	6	0.105	High — Bare/poorly coated, survivorship bias
1960–69	473,354	28	0.059	Moderate — Largest cohort (22% of network)
1940–49	140,919	6	0.043	Moderate — War-era pipe
1950–59	449,493	21	0.047	Moderate — Second largest cohort (21%)
1980–89	171,412	7	0.041	Lower — Better coatings
2000–09	196,725	8	0.041	Lower — Modern materials
1990–99	205,835	4	0.019	LOWEST — FBE coatings + effective CP

This pattern has profound implications for the modeling approach. A simple age-based regression would produce a poor fit because the 1970s vintage (age ~50) fails at 3.8× the rate of the 1990s vintage (age ~30), while older pipe from the 1950s fails at a lower rate than the 1970s. The non-monotonicity reflects technological transitions in coating materials, cathodic protection effectiveness, and installation practices across decades. This validates the need for a multivariate ML model that captures vintage-technology interactions, and demonstrates that ROSEN's ILI data (which includes actual coating condition and wall thickness) would add predictive power beyond what age alone provides.

4. Data Inventory & File Structure

4.1 Files on Disk

All project data is stored at C:\Phmsa with the following structure:

Raw Data (C:\Phmsa\annual_gt\)

15 Excel files: annual_gas_transmission_gathering_YYYY.xlsx for years 2010–2024 (100.4 MB total). Each file contains sheets for Parts A–D, F–G, H, I, J, K, L, M, N–O, P (and Q, R in later years).

Extracted CSVs (C:\Phmsa\annual_gt\extracted_csvs\)

75 CSV files extracted from the xlsx files using the custom extract_phmsa_parts.py script. Five Parts extracted per year (J, K, M, H, A–D) × 15 years = 75 files. All Part J files verified with identical schema (117 columns).

Incident Data (C:\Phmsa\)

PHMSA_Pipeline_Safety_Flagged_Incidents.zip containing Excel files with incident data from 1986 to present, organized by time period and pipeline system type.

Processed Data (C:\Phmsa\)

File	Rows	Size	Description
survival_panel_7yr.csv	34,226	2.7 MB	7-year survival panel (2017–2024) with incident matching
Survival_table_2024_prototype.txt	5,119	633 KB	Single-year prototype with covariables from Parts H, K, M

4.2 Python Environment

A dedicated virtual environment (rosen_env) has been created at C:\Phmsa\rosen_env with Python 3.12 and the following packages: openpyxl, pandas, numpy, scipy, scikit-learn, statsmodels, lifelines, lightgbm, matplotlib, seaborn, jupyter, and ipykernel. The environment is activated with: .\rosen_env\Scripts\Activate.ps1

4.3 Scripts

Script	Purpose	Input	Output
extract_phmsa_parts.py	Extract Part sheets from xlsx to CSV	annual_gt*.xlsx	extracted_csvs\GT_AR_YYYY_Part_X.csv

5. Model Design Specification

5.1 Observation Unit

Each row in the survival panel represents one (operator × state × installation_decade × report_year) combination. For example: Operator 31618 (Southern Natural Gas) in Alabama, pipe installed in the 1960s decade, observed in report year 2022. The response is the count of incidents (0, 1, 2, ...) and the exposure is the miles of pipe at risk.

5.2 Variable Definitions

Variable	Type	Source	Description
operator_id	Identifier	All Parts	PHMSA operator identification number
state	Identifier	All Parts	State where pipe is located
year	Time	Report year	Calendar year of observation
decade_bin	Categorical	Part J	Installation decade (pre1940 through 2020–29)
install_midpoint	Numeric	Derived	Midpoint year of decade (e.g., 1965 for 1960s)
age_at_obs	Numeric	Derived	year − install_midpoint (approximate pipe age)
miles_at_risk	Exposure	Part J	Miles of transmission pipe (Poisson offset)
n_incidents	Response	Incidents	Count of all-cause incidents
n_corrosion	Response	Incidents	Count of corrosion incidents
n_material	Response	Incidents	Count of material/weld incidents
event	Binary	Derived	1 if n_incidents > 0
event_corrosion	Binary	Derived	1 if n_corrosion > 0

Covariables (to be joined from Parts D, H, K, M in the modeling phase):

Covariable	Source	Type	Description
pct_diam_small	Part H	Continuous [0,1]	Fraction of miles with diameter ≤8"
pct_diam_24	Part H	Continuous [0,1]	Fraction of miles with diameter 20–24"
pct_diam_30	Part H	Continuous [0,1]	Fraction of miles with diameter 26–30"
pct_smys_high	Part K	Continuous [0,1]	Fraction of miles at >60% SMYS
pct_class1	Part K	Continuous [0,1]	Fraction of miles in Class 1 (rural)
n_ext_corrosion	Part M	Count	Leak/repair events from

			external corrosion
n_int_corrosion	Part M	Count	Leak/repair events from internal corrosion
total_trans_miles	Part D	Continuous	Total transmission miles (operator size proxy)

5.3 Modeling Approach (Three Phases)

Phase 1: Poisson GLM (MVP)

The baseline model uses Poisson regression with log(miles_at_risk) as an offset, which is equivalent to modeling the incident rate per mile-year. The model equation is: $\log(E[\text{incidents}]) = \log(\text{miles}) + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{vintage} + \beta_3 \cdot \text{diameter\_mix} + \beta_4 \cdot \text{smys\_class} + \beta_5 \cdot \text{leak\_history} + \beta_6 \cdot \text{state}$. This approach is interpretable, computationally efficient, and provides baseline hazard rate estimates with confidence intervals.

Phase 2: Random Survival Forest

A non-parametric ensemble model that captures nonlinear interactions between covariables (e.g., coating era × age, diameter × %SMYS). Provides variable importance rankings and partial dependence plots for interpretability. Expected to improve on the Poisson GLM by capturing the non-monotonic vintage effects without manual feature engineering.

Phase 3: DeepSurv / Competing Risks

A neural network extension of the Cox proportional hazards model that handles competing risks (corrosion vs material failure vs excavation damage). State-of-the-art for complex survival data. Will integrate the Mendeley ILI dataset to demonstrate value of pipe-level inspection data.

5.4 Statistical Power

Panel	Observations	All Events	Corrosion Events	Max Predictors (10 EPV rule)
Current (7 years)	34,226	657	106	10 corrosion / 65 all-cause
Full (15 years)	~72,000	~1,400	~210	21 corrosion / 140 all-cause

The 10 Events Per Variable (EPV) rule of thumb for regression indicates that the 7-year panel supports up to 10 covariables for corrosion-specific models and 65 for all-cause models. The full 15-year panel approximately doubles these limits.

6. Next Steps & Roadmap

6.1 Immediate (Data Completion)

- Expand survival panel from 7 years to 15 years using extracted Part J CSVs (2010–2016 + 2019 now available)
- Join covariables from Parts D, H, K, and M to the survival panel
- Validate incident matching rates for 2010–2016 years

6.2 Phase 1: Poisson GLM (1–2 weeks)

- Fit Poisson GLM with age + vintage + material + diameter covariables
- Compute cause-specific hazard rates (corrosion, material, all-cause)
- Validate against held-out year (2024) and temporal cross-validation
- Generate operator-level risk scores and risk ranking

6.3 Phase 2: Advanced Models (2–4 weeks)

- Random Survival Forest with full feature matrix and hyperparameter tuning
- Integrate Mendeley ILI dataset for pipe-level metal loss validation
- ERA5 environmental features (temperature, precipitation via lat/lon)
- Variable importance analysis and partial dependence plots

6.4 Phase 3: Deliverables

- Interactive risk dashboard (React/Plotly) for operator-level risk visualization
- Technical paper draft suitable for conference submission (e.g., ASME IPC, Pipeline Pigging & Integrity Management)
- ROSEN pitch deck with model outputs demonstrating ILI data value proposition

7. References & Data Sources

PHMSA Source Data Portal: https://www.phmsa.dot.gov/data-and-statistics/pipeline/source-data

PHMSA Annual Report Data: https://www.phmsa.dot.gov/data-and-statistics/pipeline/gas-distribution-gas-gathering-gas-transmission-hazardous-liquids

PHMSA Flagged Incident Files: https://www.phmsa.dot.gov/data-and-statistics/pipeline/source-data (Flagged Files section)

Mendeley ILI Data: https://data.mendeley.com/ (search: probabilistic degradation models pipeline)

49 CFR Part 191 — Transportation of Natural and Other Gas by Pipeline; Annual Reports, Incident Reports, and Safety-Related Condition Reports

49 CFR Part 192 — Transportation of Natural and Other Gas by Pipeline: Minimum Federal Safety Standards

API 5L — Specification for Line Pipe

ASME B31.8 — Gas Transmission and Distribution Piping Systems

López de Prado, M. (2018). Advances in Financial Machine Learning. Wiley. (Methodology reference for anti-data-leakage and walk-forward validation frameworks)