AI-POWERED SOLUTIONS FOR MISSING DATA IN PIPELINE RISK ASSESSMENTS

Syed Jehanzeb Adeel Haider

Enbridge Energy Inc., Houston, USA

ABSTRACT

The integration of Artificial Intelligence (AI) and Machine Learning (ML) into the pipeline sector of the oil and gas industry has demonstrated considerable potential, especially in overcoming the difficulties associated with incomplete data. This paper explores the application of AI in supplementing missing data for risk evaluations, particularly in scenarios where safety is a critical concern. Potential pitfalls and risks associated with relying solely on AI-generated data are analytically discussed and illustrated in this paper. Through a detailed process flow, this paper also suggests strategies to balance AI reliance with real data acquisition, emphasizing the importance of consequence analysis, cost-benefit considerations, and a hybrid approach to ensure the safety and reliability of operations across the pipeline and broader oil and gas industry in an efficient way.

KEYWORDS

Artificial Intelligence (AI), Machine Learning (ML), Risk Assessment, Pipeline Safety, ALARP

1. INTRODUCTION

Safety-critical risk assessments in the oil and gas industry depend heavily on accurate and complete data. Many vintage pipelines and other infrastructure elements are often lacking comprehensive key attribute information (e.g., pipe grade) which creates a major problem. Traditional methods to determine missing data are typically expensive and technologically demanding. This paper discusses how AI and ML can be employed to fill these data gaps and highlights potential errors and their implications. A structured decision-making process is presented to provide guidelines on the use of AI and the need of real data acquisition in risk assessments, applicable across various scenarios in the oil and gas industry [1].

2. THE CHALLENGE OF MISSING DATA IN THE OIL AND GAS INDUSTRY

In many segments of the oil and gas industry, operators face the issue of missing data. Determining this missing data using current technology is typically costly and complex. In scenarios where data is scarce, AI and ML present a viable solution by predicting missing values based on available data. However, this approach comes with risks, especially when the data is used in Safety Critical risk assessments. [2].

Bibhu Dash et al:COMSCI, IOTCB, AIAD, EDU, MLDA, EEEN - 2024
pp. 149-155, 2024. IJCI – 2024 DOI:10.5121/ijci.2024.130412

3. AI TECHNIQUES FOR ADDRESSING MISSING DATA

AI techniques such as supervised learning and unsupervised learning can be applied to predict missing data. However, the accuracy of these models heavily depends on the quality and relevance of the training data [3].

3.1. Supervised Learning

This method involves training a model on a labeled dataset, where the target outcome is known. The model learns to predict the outcome for new data based on patterns identified during training [3], [5].

- **Linear Regression**: Predicts continuous values, useful for estimating missing data based on linear relationships.
- **Decision Trees**: Captures non-linear relationships and interactions between different features.
- **Random Forests**: An ensemble method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.

3.2. Unsupervised Learning

This approach deals with unlabeled data. The model tries to find hidden patterns or intrinsic structures in the input data without explicit instructions on what to predict.

- **K-Means Clustering**: Groups data into clusters based on feature similarity, which can help infer missing data from similar clusters [4].

4. BALANCING INNOVATION WITH PRUDENCE

While AI and ML offer innovative approaches to filling data gaps, their limitations must be acknowledged. Ensuring the integrity and quality of training data is crucial. AI models should be validated against known data points and reviewed by field experts. Transparency about the limitations of AI techniques used is essential, and a logical structured approach should be utilized when making decisions based on AI-generated data.

Non-conservative interpretations of AI-generated data can lead to an underestimation of risks, potentially overlooking high-consequence scenarios. It is vital to validate AI predictions and maintain transparency about the AI's limitations to ensure the responsible use of AI in risk assessments.

A hypothetical example of a safety critical risk scenario where data generated by AI / ML model might lead to an inaccurate and non-conservative risk evaluation is discussed in the following section.

5. HYPOTHETICAL EXAMPLE. PIPE GRADE DATA IN A VINTAGE PIPELINE

Pipeline regulations generally recommend a cautious approach, but sometimes operators are allowed to use a rational and less conservative method for hard-to-acquire missing data. A hypothetical case scenario is studied in this paper to identify potential issues that could occur when using AI to determine the missing safety critical data in pipeline risk assessments. Consider a vintage gas pipeline where substantial pipe grade data is unavailable. After residential construction in the area, the pipeline section is reclassified as High Consequence Area (HCA). The pipeline operator, aiming to minimize costs, employs an ML model to predict the missing pipe grade data. The model, using a conservative HCA-compatible pipe grade value, overlooks the original non-HCA classification and less conservative pipeline grade. This misclassification results in non-conservative risk assessments, potentially leading to unsafe conditions [1]. This example illustrates a broader issue within the oil and gas industry where missing data can lead to significant safety and operational risks if not handled correctly. The strategies and processes discussed here can be applied to various other scenarios such as equipment maintenance schedules, safety inspections, and environmental monitoring data gaps.

5.1. Potential Pitfalls and Non-Conservative Outcomes

Errors in AI predictions can lead to non-conservative risk assessments. In the given example, the ML model's conservative assumptions for the HCA classification led to an underestimation of risks. Such errors can have dangerous outcomes, emphasizing the need for careful consideration when relying on AI-generated data [1].

To illustrate the two scenarios (AI-generated data risk vs. real data risk), consider the following dataset for a vintage gas pipeline:

Table 1. Risk results for hypothetical vintage pipeline with real and AI generated data.

Segment ID	**HCA Status at the time of construction**	**HCA Status - Current**	**AI Predicted Pipe Grade**	**Real Pipe Grade**	**Risk Result with AI Data**	**Risk Result with Real Data**
1	No	No	X42	X42	$10,000	$10,000
2	No	No	X42	X42	$10,000	$10,000
3	Yes	Yes	X70	X70	$20,000	$20,000
4	Yes	Yes	X70	X70	$20,000	$20,000
5	No	No	X42	X42	$10,000	$10,000
6	No	No	X42	X42	$10,000	$10,000
7	No	Yes	X70	X42	$20,000	$40,000
8	No	Yes	X70	X42	$20,000	$40,000

This example shows that segments 7 and 8 changed to HCA status after residential construction happened near the pipeline, which was built earlier. AI assigned a higher-grade prediction to these segments, using the logic that HCA requires higher pipe grade, even though the lower pipe grade was used when the location was not HCA at the time of construction. This resulted in non-conservative assessment with the data generated by AI. The real data shows a lower grade (X42) than the AI-predicted grade (X70), highlighting the discrepancy and potential risk with AI generated data as shown in Figure 1.

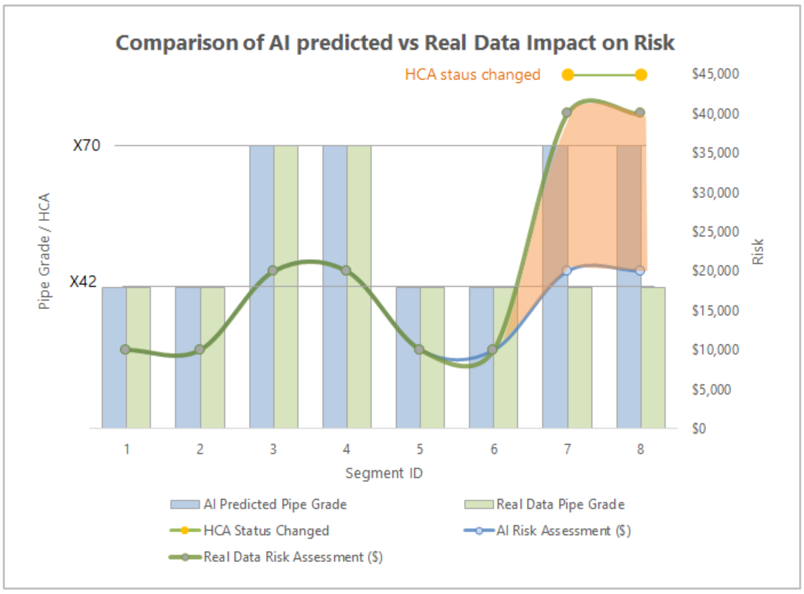

Figure 1. AI generated vs real data impact on risk for hypothetical pipeline example.

6. Strategies for Balancing AI and Real Data Acquisition

To reduce the potential hazards of relying on uncertain or less cautious data for risk assessments especially in safety critical situations, operators should follow a systematic method and decision-making process. A suggested process that could provide a useful framework is presented in the process flow diagram in Figure 2.

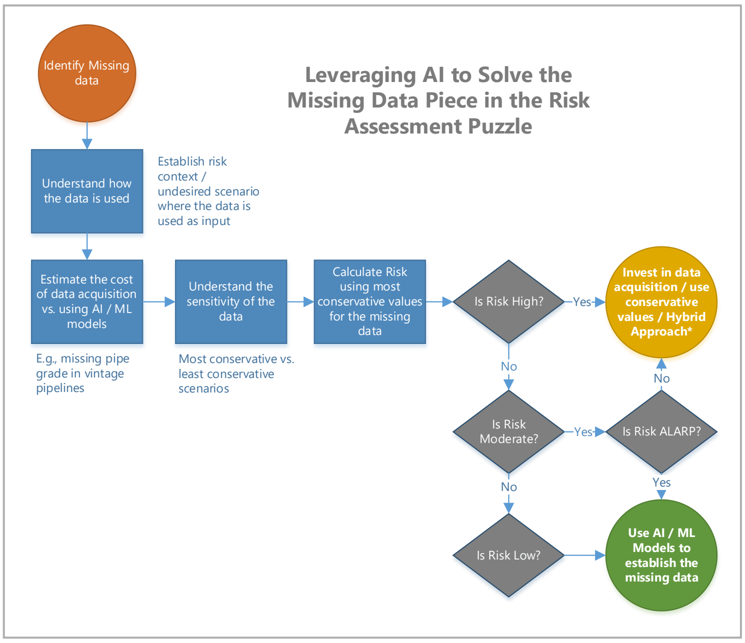

Figure 2. Process diagram suggesting steps on risk-based approach on how to leverage AI for missing data

6.1. Process Steps

- **Identifying Missing Data:** This is the foundational step in establishing the risk context. It requires a thorough cataloging of all required data, determining the exact locations where data is missing, and assessing the impact of these gaps on the overall risk assessment. This stage lays the foundation for an in-depth understanding of the data and its consequences for safety and making risk informed decisions.
- **Understanding How the Data is Used**:This involves mapping the data to its corresponding risk model and evaluating its relevance and significance in the overall risk calculation. This understanding is vital to determine the potential consequences of data absence on the risk assessment process.
- **Estimating the Cost of Real Data Acquisition versus Using AI/ML Models:** This step is critical for economic and strategic consideration. This involves a cost-benefit analysis that compares the expenses and benefits of traditional data acquisition against the predictive capabilities of AI/ML models. Additionally, the timeframe for each method is evaluated to ensure alignment with project deadlines and objectives.
- **Understanding the Sensitivity of the Data:** This is a pivotal step that involves conducting sensitivity analyses to understand how variations in data can affect risk outcomes.This analysis rigorously compares the effects of the most conservative educated assumptions on data against the least conservative ones, determining how each could potentially influence the final risk evaluation.
- **Calculating Risk Using Conservative Values for the Missing Data:** This step ensures that the risk levels are determined based on the safest possible estimates. This involves applying conservative assumptions to the missing data, quantifying the risk levels, and having the calculations reviewed and validated by experts. This step is essential for confirming the accuracy and reliability of the risk assessments and for maintaining high standards of safety.

6.2. Decision Points Explained

- **High Risk**: If the calculated risk is high, it is crucial to invest in data acquisition, use conservative values, or adopt a hybrid approach. This ensures that the critical areas with high potential for failure are addressed with the highest accuracy.
- **Moderate Risk**: For moderate risk levels, it is essential to evaluate whether the risk is ALARP (As Low As Reasonably Practicable). If the risk reduction measures are reasonably practicable and not disproportionately costly, they should be implemented to ensure safety.
- **ALARP Determination**: If the risk is determined to be ALARP, AI/ML models may be used to establish the missing data. This balances cost and safety by applying advanced AI techniques to manage acceptable risk levels.
- **Low Risk**: When the risk is low, AI/ML models can be effectively used to establish the missing data without significant investment in real data acquisition, ensuring operational efficiency.

Decision on whether to invest in data acquisition, use conservative values, or to use a hybrid approach depends on the risk levels and the cost benefit analysis of the cost of data acquisition and potential risk benefit. Potential Risk benefit is the difference between the highest and the lowest conservative data values.

6.3. Alarp (As Low As Reasonably Practicable): Ensuring Acceptable Risk Levels

The ALARP principle is used to decide if the risk is as low as it can reasonably be. The methodology to determine if the risk is ALARP involves:

- **Comparing Costs and Benefits**: Weighing the cost and effort required to reduce the risk further against the potential benefits.
- **Proportional Measures**: Ensuring that risk reduction measures are proportionate to the risk level.
- **Adopting Additional Measures**: Adding more safety measures if they are reasonably practicable, meaning the costs don't outweigh the benefits [7].

6.4. Hybrid Approach: Merging AI with Strategic Data Acquisition

The process implies that a hybrid approach may be a better strategy than using the most conservative assumptions on data or spending a lot of resources on extensive data acquisition for unacceptable risk situations. Therefore, it is important to know what a hybrid approach means in this context. The hybrid approach involves strategically acquiring critical data points to support and improve the AI/ML model's predictions for the missing data. This approach is particularly useful when the cost of acquiring comprehensive real data is unreasonable. The process includes:

- **Identifying Critical Data Points**: Focus on the most impactful data for risk assessments.
- **Using AI/ML Models**: Fill in less critical data gaps with AI predictions.
- **Continuous Validation**: Regularly validate and update AI predictions with newly acquired real data to enhance model accuracy and reliability.

This strategy minimizes risk when it comes to safety-critical data by ensuring that the most impactful and potentially hazardous data gaps are filled with highly reliable real data, while less critical gaps are managed through AI, which is continuously validated and improved.

7. Creating Guidelines For AI Application

To navigate the complexities of Artificial Intelligence, we must adhere to a set of stringent guidelines that ensure the safety and accuracy of AI models. The steps below describe the essential measures needed to preserve the reliability of AI models.

- **Validation and Verification**: Implement strict validation and verification processes for AI models. This includes cross-referencing AI-generated data with available real data and using multiple models to compare results.
- **Conservative Safety Margins**: Apply conservative safety margins to AI predictions to account for uncertainties and potential errors in the model.
- **Continuous Monitoring and Updating**: Regularly monitor and update AI models to ensure they remain accurate and relevant as new data becomes available.
- **Clear Documentation:** Maintain clear documentation of the AI models used, including their assumptions, limitations, and the training datasets.
- **Risk-Based Approach**: Adopt a risk-based approach to determine the balance between AI reliance and real data acquisition, focusing on the potential consequences of failure in different segments. A thorough cost-benefit analysis should be conducted to justify the investment in real data acquisition versus relying on AI predictions [6].

8. CONCLUSION

AI and ML offer valuable tools for addressing missing data in pipeline risk assessments and broader applications in the oil and gas industry. However, their application must be balanced with real data acquisition to ensure safety and reliability. By establishing robust guidelines and cautious implementation, the oil and gas industry can leverage AI's benefits while mitigating potential risks. The decision to rely on AI or invest in real data acquisition should be based on a detailed consequence and risk analysis as well as cost-benefit considerations. When the stakes are high, such as in HCAs or critical infrastructure, real data acquisition is highly recommended. For less critical systems, AI can provide a cost-effective and efficient alternative, however it must be applied with caution and validated rigorously.

REFERENCES

[1]. "Pipeline Safety and Integrity Management," Enbridge Inc., 2021. [Online]. Available: Enbridge Pipeline Safety.

[2]. "Pipeline Integrity Management," Department of Transportation, Pipeline and Hazardous Materials Safety Administration, 2019. [Online]. Available: PHMSA Pipeline Integrity Management.

[3]. Quinlan, J.R., "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.

[4]. MacQueen, J., "Some Methods for Classification and Analysis of Multivariate Observations," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281-297, 1967.

[5]. Breiman, L., "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[6]. Bowden, R.O., Lane, M.P., and Martin, J.D., "Integrated Risk Management in Oil and Gas Pipelines," Journal of Risk and Reliability, vol. 15, no. 4, pp. 231-246, 2001.

[7]. "Guidance on ALARP Decisions in the Management of Major Hazards," Health and Safety Executive (HSE), UK, 2009. [Online]. Available: HSE ALARP Guidance

AUTHOR

Syed Jehanzeb Adeel Haider, Registered Professional Engineer in Alberta, Canada. Master of Engineering from University of Alberta, Canada. Experienced engineering professional with nearly two decades of dedicated service in the oil and gas sector, specializing in risk assessment, reliability engineering, and integrity management. Known for leading project teams with a focus on developing innovative risk management strategies and optimizing processes, I bring a thoughtful approach to mentorship and a commitment to effective communication with stakeholders at all levels.