**CAAP Quarterly Report**

**April 5th 2023**

*Project Name: Pipeline Risk Management Using Artificial Intelligence-Enabled Modeling and Decision Making*

*Contract Number: 693JK32150001CAAP*

*Prime University: Rutgers University*

*Prepared by: Bingyan Cui (PhD student), Emad Farahani (PhD student), Dr. Hao Wang (PI), Dr. Qindan Huang (Co-PI)*

*Reporting Period: 1/1/2023 – 3/31/2023*

**Project Activities for Reporting Period:**

*Task 1 Literature Review (Completed)*

*Task 2 Data Collection from Industry Partners (Completed)*

*Task 3 Data-Driven Probabilistic Modeling of Pipeline Defects*

***Predicting Defect Growth Using Bayesian Neural Network***

*BNN Model Development with Zone-Based Data*

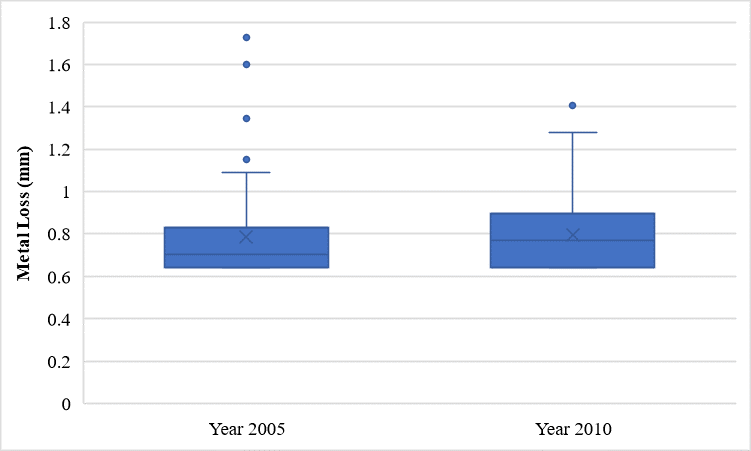
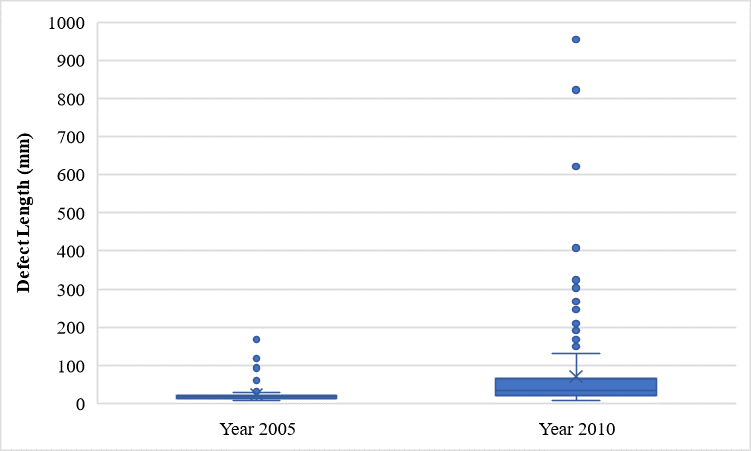
In the dataset from the 112-km pipeline in Mexico, the soil properties are measured per zone but there are various measurement values of pipeline defects at each zone, which cannot be explained by the variation of soil properties. One solution is to use zone-based data in the prediction model. Although this will reduce the size of dataset, the model accuracy is expected to be improved by reducing the uncertainty of input data itself.

The ILI data were further processed to generate the zone-based dataset. One pair of corrosion defects from two inspections that matched with each other at the nearest location and showed the increasing trend were extracted from each zone to generate the dataset. Figure 1 shows the scatter plot of corrosion defects along pipeline length in the zone-based dataset. Figure 2 and 3 illustrate the boxplots and distribution frequency plots of defects in the zone-based dataset, respectively, for metal loss and defect length. As compared to the processed dataset based on all defect measurements, the variance of defects was reduced by removing the multiple data points within the same zone.

(a)

(b)

Figure 1 Scatter plot of corrosion defects along pipeline length for the zone-based dataset: (a) metal loss; and (b) defect length

1. (b)

Figure 2 Boxplot for the zone-based dataset: (a) metal loss; and (b) defect length

1. (b)

Figure 3 Distribution frequency plot for the zone-based dataset: (a) metal loss; and (b) defect length

The zone-based datasets were used to establish BNN models for the prediction of metal loss and defect length. The dataset was divided into the training set (80%) and test set (20%) for cross-validation. In this study, the inputs included the soil parameters (Eh, resistivity, pH, concentration, soil moisture, and soil type), age of pipe segment (after initial installation or replacement), elevation, and wall thickness. BNN models were developed for metal loss and defect length separately, and the coefficients of determination (R2) for model accuracy (based on test set) were summarized in Table 1.

Table 1 Accuracy of BNN models for prediction of pipe defects (zone-based dataset)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Soil Property Input** | **Defect Type** | **BNN** | **Dataset Size** |
| Zone-Based Dataset | Original | Metal Loss | 0.64 | 380 |
| Defect Length | 0.94 | 326 |

Figure 4 shows the measured vs. predicted corrosion defects with 95% conference interval with the BNN models developed using the zone-based dataset. The blue points represented the measured values, the orange points stand for the mean value of predicted results, and the orange range corresponded to the confidence interval which was expected to cover 95% of predictions from the view of statistics.

(a)

(b)

Figure 4 Measured vs. predicted defects with 95% conference intreval by BNN models developed using the zone-based dataset: (a) metal loss; and (c) defect length

*Shapley Additive Explanation of BNN Model*

Sophisticated machine learning algorithms can provide accurate predictions, but it is challenging to interpret the model. Shapley Additive Explanation (SHAP), a unified and game theoretic tool, is applied to explain the output of machine learning model and increase transparency based on the game theoretically optimal Shapley values. In game theory, the principle of SHAP value is to evaluate the contributions from each player to the game result separately, while keeping the sum of contributions being equal to the final game result. Thus, in model interpretation, SHAP is used to measure each feature’s contributions to the final prediction of model by assigning a SHAP value to each feature. A higher average SHAP value indicates a more important parameter. SHAP values interpret the process of deriving the final model output by starting from the base value that would be predicted if none of the features was known, after which SHAP values are incrementally calculated conditioned on one feature at a time. The feature’s contribution is determined by accumulating all the feature combinations. When the addition of a feature increases the output value, it has a positive SHAP value.

In this study, global interpretation of derived model was conducted through calculating the SHAP values, and the positive and negative correlations between variables and prediction were illustrated by the SHAP value plots. SHAP values were calculated for the BNN models of metal loss and defect length to analyze the effects of parameters on the output. SHAP values were estimated for the BNN models developed using the zone-based dataset for the propagation of metal loss and defect length.

Figure 5 shows the variable importance plots of SHAP values. The descending order of parameters indicates the rank of variable importance on the output from high to low, and the color of the points indicates the magnitude of the feature for each observation with blue standing for lower values and red standing for higher values. Soil moisture showed relatively higher importance on corrosion defects as compared to other influencing variables. It was suggested that soil moisture had positive effects on metal loss and defect length in general.

Chart

Description automatically generated with medium confidenceChart, scatter chart

Description automatically generated

(a) (b)

Figure 5 Variable importance plot of BNN models using the zone-based dataset: (a) metal loss; and (b) defect length

***Predicting Defect Growth using Probabilistic Power Law Model***

As mentioned in the previous report, a power-law function of time model formation is adopted for corrosion growth as shown in Eq. 1 below.

|  |  |
| --- | --- |
|  | (1) |
|  | (1a) |
|  | (1b) |

Where, *m* = types of defect quantity (e.g., *m* = *D* for the maximum defect depth and *m* = *L* for the maximum defect length), *Ym*= defect quantity (e.g., maximum defect depth or defect length) at a time instant *t*,**θ** = unknown model parameters, *fi* and *hj* = influencing environmental variables.

The 112 km pipeline is divided into 56 segments with 2 km length for each segment. Then linear regression is conducted by treating the resulted *C*1,*k* or *C*2,*k* as the response and the mean values of the soil parameters for each segment as the predictors. The results indicate that the soil moisture, *M*, and soil sulfate level, *SO*4, have the least *p*-value in the developed linear models for *C*1,*k* and *C*2,*k*, respectively. Thus, Eq. (1a) and (1b) become Eq. 2 as follows.

|  |  |
| --- | --- |
|  | (2a) |
|  | (2b) |

By incorporating the soil properties at the location of each defect, the unknown model parameters are assessed based on all the defect data of the whole pipeline using MCMC.

The predicted corrosion depth and length and the associated actual measured quantities obtained through ILI are compared in Figure 6. It is found that the prediction accuracy has been significantly increased when comparing the results in the previous report where only soil moisture was incorporated in the model. This result shows that using our approach to select soil properties is effective. However, more datasets should be used to continue testing and improving this approach.

|  |  |
| --- | --- |
|  |  |
| (a) | (b) |
| Figure 6 Comparison between predicted versus measured defect dimension: (a) depth and (b) length | |

***Summary***

It is noted that the developed models are applicable for a wide range of pipe defects, such as corrosion, cracks, mechanical damages, as well as interactive threats. Based on the ILI data from the industry partner, external corrosion is the dominant one that shows a growth trend.  Thus, the prediction models for external corrosion were developed and evaluated for accuracy.

Based on the results of both prediction models, the general trend of defect growth can be captured with the average corrosion depth and length. However, large variations were found in the prediction. This is probably due to the fact that the ILI data are only available for two inspections. The model accuracy would increase as more inspection data become available as model inputs. Therefore, the research team will search the available multiple-year defect data in the literature for analysis, in addition to keep refining the developed models.

**Project Activities with Cost Share Partners:**

Cost share is provided by Rutgers University and Marquette University during this quarterly period as budgeted in the proposal.

**Project Activities with External Partners:**

N/A

**Potential Project Risks:**

N/A

**Future Project Work:**

Considering the importance of accurately predicting defect growth trend for repair scheduling, the research team will continue working on Task 3 Data-Driven Probabilistic Modeling of Defects and delay the start of Task 4 Quantification of Probability of Failure.

1. Considering there are only two inspection data available, the corrosion rate can be calculated and used as the dependent variable to be predicted. In this case, BNN model will be used to predict how soil environment and other variables affect the relative growth of corrosion defect between two inspections. In addition, assumptions will be made to consider random generation of defects in the BNN model.
2. The prediction accuracy of probabilistic power-law models will be further improved by incorporating more soil properties or dividing the pipe length into different zone segments where the prediction model at each zone will have different model parameters.
3. The availability of multi-year inspection data will be further checked in the literature to see if it is possible to demonstrate the developed approach for different datasets and compare the accuracy.

**Potential Impacts to Pipeline Safety:**

The ILI data will be used to develop probabilistic growth models of pipeline corrosion defects, which can aid pipeline operators better predict failure risk and make repair decisions.