

Quarterly Report – Public Page

Date of Report: 5th Quarterly Report-December 31st, 2020

Contract Number: 693JK31910018POTA

Prepared for: DOT PHMSA

Project Title: Mapping Indication Severity Using Bayesian Machine Learning from Indirect Inspection Data into Corrosion Severity for Decision-Making in Pipeline Maintenance

Prepared by: TEES (Texas A&M Engineering Experiment Station) and University of Dayton

Contact Information: Homero Castaneda, hcastaneda@tamu.edu, 979 458 9844.

For quarterly period ending: December 30th, 2020

1: Items Completed During this Quarterly Period:

The following activities have been completed:

<i>Item</i>	<i>Task</i>	<i>Activity/Deliverable</i>	<i>Title</i>	<i>Federal Cost</i>	<i>Cost Share</i>
9	1,2,3	5 th Quarterly Report	5th Quarterly Report	4,000.00	0.00
4	3	Three unsupervised learning strategies (k-means, Gaussian Mixture Model, and Hidden Markov Random Field) for soil corrosivity clustering.	Unsupervised learning strategies	20,000.00	3,000.00
6	3	Two supervised learning strategies (Support Vector Machine, Relevance Vector Machine) for defect type classification	Two supervised learning strategies	20,000.00	3,000.00

The title of the table is based on the file Technical and Deliverable Payable Milestone

2: Items Not-Completed During this Quarterly Period:

Task number 2, extract basic corrosion model parameters started during previous quarters. Part of Task 2 will be cover in the following reports. During the experimental and analysis activities we found information that influence the time extension. The following activities will be ready in latter reports based on the Technical and Deliverable Payable Milestone

<i>Item #</i>	<i>Task #</i>	<i>Activity/Deliverable</i>	<i>Title</i>	<i>Federal Cost</i>	<i>Cost Share</i>
8	2	Extract basic corrosion model and embed into the previously developed stochastic corrosion rate model framework	Experiments and analyses to bridge gaps in prior knowledge	24,000.00	0.00
10	4	Bayesian regression for corrosion rate model calibration	Bayesian Regression analysis	30,000.00	3,000.00

3: Project Financial Tracking during this Quarterly Period:

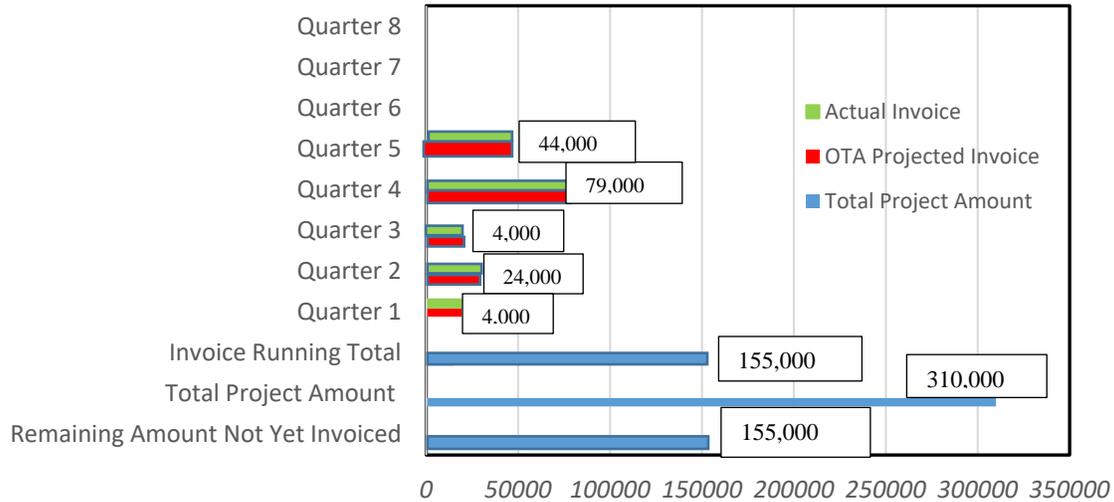
The table has been updated based on the deliverables and corrected attachment No5 Technical and Deliverable Payable Milestone.

Date of Report: 5th Quarterly Report –

**Prepared by Homero Castaneda, TEES
Hui Wang, U. Dayton**

Page 1

Quarterly Payable Milestones/Invoices 693JK31910018POTA



4: Project Technical Status –

The following tasks are included in the project:

- **Task 2: Experiments and analyses to bridge gaps in prior knowledge**
- **Task 3: Bayesian machine learning to bridge gaps in uncertainty quantification.**
- **Task 4: Finalize and evaluate/validate the model.**

During the fifth quarter, the team members from Texas A&M University (TAMU) and the University of Dayton (UD) had biweekly meetings and an internal Workshop to discuss, characterize, analyze and understand the field/ laboratory results and how we extract information to generate and develop new correlations and trends for the prioritization of the survey and inspection tools.

The team organized an internal Workshop entitled “Data Driven Integrity Pipeline Management”. The outcomes of the workshop will help the PhD students in both TAMU and UD teams with different knowledge backgrounds to understand the mathematical tools used to characterize, classify and quantify the parameters sensed in the field and measured in the laboratory with respect to severity due to corrosion.

This report includes the activities based on the proposed schedule for tasks 2,3 and 4 as follows:

Task 2: Experiments and analyses to bridge gaps in prior knowledge

During this quarter, we performed the proposed experimental matrix (see Table 1) to develop new correlations with parameters founded in the field and also in laboratory conditions. The results founded in previous quarter helped to initiate new extraction of results that added more data to start to build new relationships and expression under control conditions (laboratory). The previous parameters included field simulations in the laboratory conditions, this quarter the reproducibility, cathodic protection potential, different pH were included with different coating failure conditions. The coating used include thicker and more homogeneous layer (the tolerance was less than 1 mils in thickness). Three different conditions are considered for the coating

anomalies and conditions (i.e., intact) and coded (or related) to deterministic and probabilistic modeling by following the corrosion mechanism. The results of the experimental testing have been revealed the difference between intact (or no defect), active surface due to a coating defect and passive surface with corrosion products layer induced and formed on the defect area.

Experimental set up

The performed set of laboratory experiments include the effects of pH and the metallic surface condition in the presence of holidays (specifically intact, active and passive state) under different levels of cathodic protection. The experimental design performed is presented in Table 1. Buffer solution (With a buffer solution, **pH = 10.6 NaHCO₃/Na₂CO₃** , **pH = 3.6 sodium acetate/acetic acid**) used for the experiments. The passive holiday can be realized by external anodic current via potentiostat (Gamry, The Interface 600plus™). NS4 solution with composition (unit: g/L) of KCl: 0.122, NaHCO₃: 0.483, CaCl₂.2H₂O: 0.093 and MgSO₄.7H₂O: 0.131 is used to simulate soil conditions.

Sequence of non-destructive techniques electrochemical techniques for meaningful parameters that correlates the results or parameters in the field. The methods include: the evolution of *on* and *off* potential for cathodic protection conditions (under-protection and over protection), The potential will be included for DC and AC methods to characterize the elements of the electrochemical cell.

The experimental set up includes the following parameters:

Sample	Soil Composition	Coatings Thickness (coal tar 300B)	Cathodic Protection (mV vs. CuSO ₄ /Cu)	Severity based on active-passive concept	pH
API X52	NS4	25 mils	-850	Active Holiday	4
API X52	NS4	25 mils	-850	Active Holiday	7
API X52	NS4	25 mils	-850	Active Holiday	10
API X52	NS4	25 mils	-850	Passive Holiday	4
API X52	NS4	25 mils	-850	Passive Holiday	7
API X52	NS4	25 mils	-850	Passive Holiday	10
API X52	NS4	25 mils	-1000	Active Holiday	4
API X52	NS4	25 mils	-1000	Active Holiday	7
API X52	NS4	25 mils	-1000	Active Holiday	10
API X52	NS4	25 mils	-1000	Passive Holiday	4
API X52	NS4	25 mils	-1000	Passive Holiday	7
API X52	NS4	25 mils	-1000	Passive Holiday	10

Table 1 Experimental design matrix for electrochemical measurements

The area of the defect included a size is a square with length of 40 cm. The defect size includes and area of 0.5 cm * 0.5 cm. Figure 1 shows the different potential conditions, the overprotection magnitude and protection magnitude present different potential decay response. The passive state of the holiday surface shows different decay and also recovery potential response. The cathodic protection potential influences the severity response of the system.

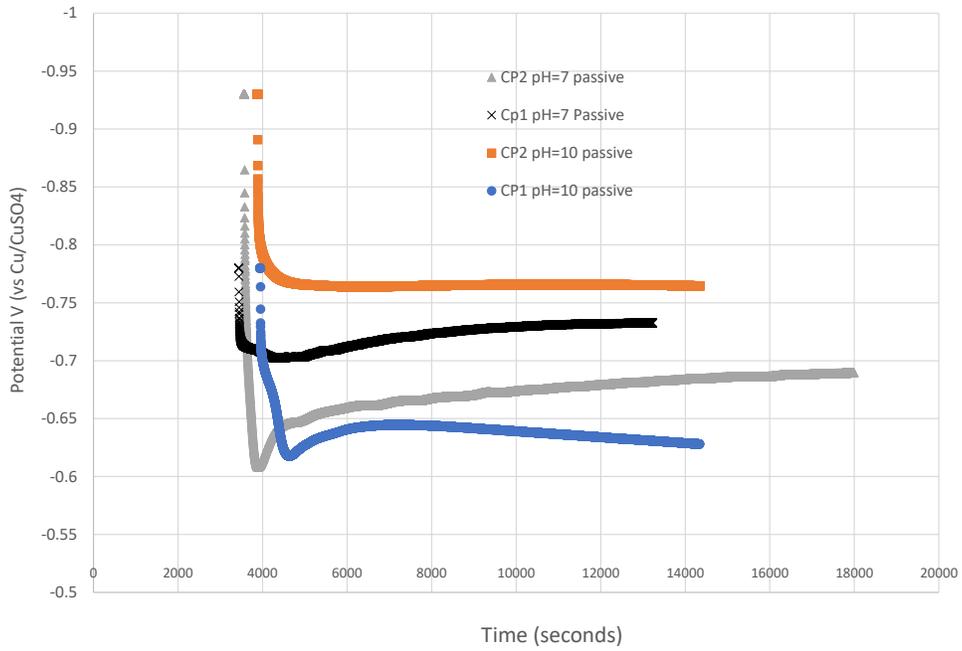


Figure 1. Potential on-off for severity classification for different Cathodic protection conditions.

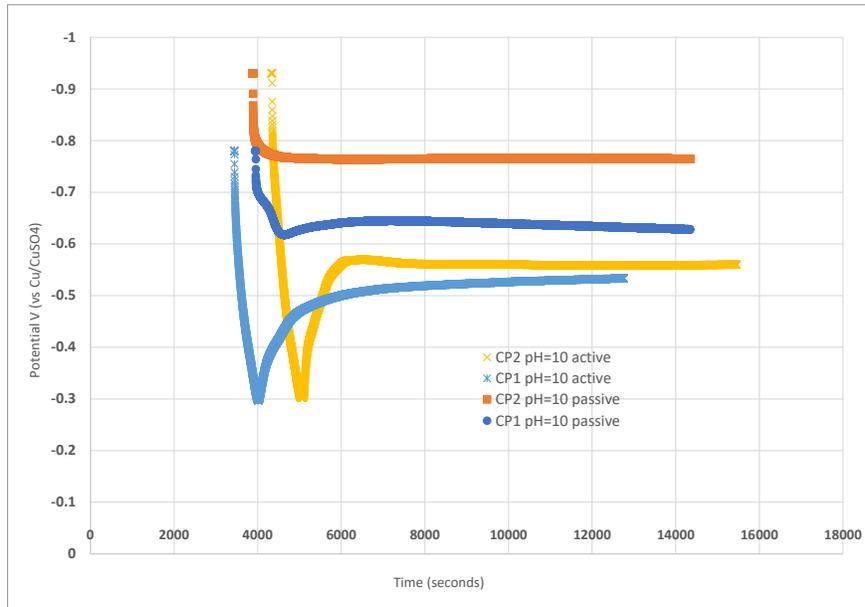


Figure 2. Potential on-off for severity classification for different pH conditions.

Potential magnitude to distinguish active and passive states.

The laboratory conditions can be marked in the E-pH diagram (see Figure 3), the severity level can be a combination of the three regions shown in the diagram with different passivity conditions. The stable phases for the iron diagram are Fe, Fe²⁺, Fe₃O₄ and Fe₂O₃.

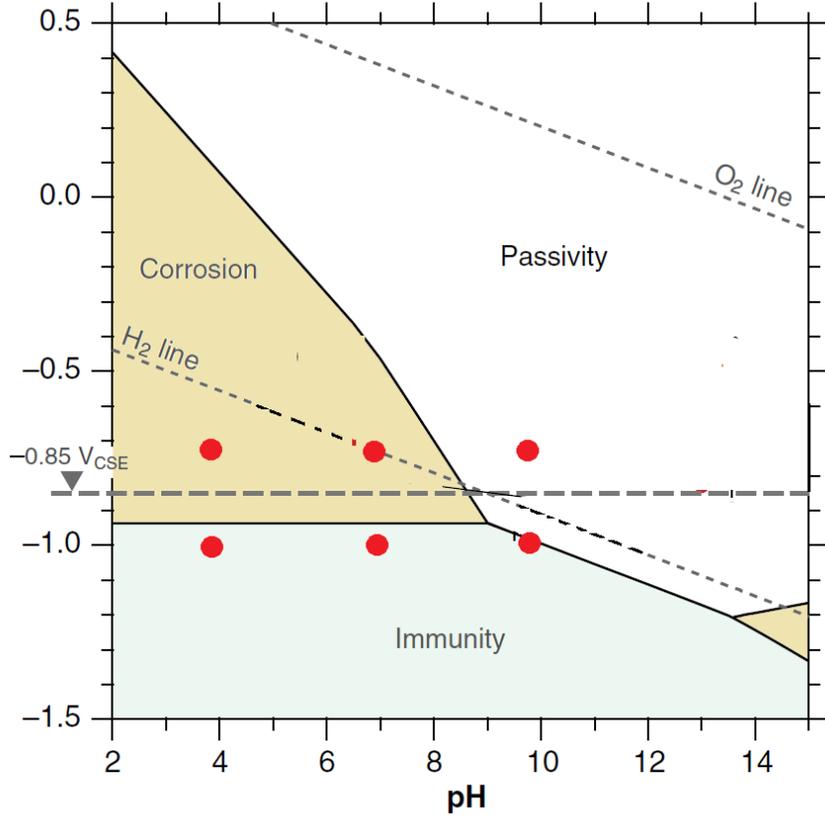


Figure 3. E-pH diagram for Iron and the stable regions for the experimental testing

Different levels for severity

During this quarter the field data was examined based on the master table originated initially. The severity can be divided based on the thermodynamics of the iron/soil interface. The severity levels can be named as follows:

Severity level Ranking	Surface condition	Equilibrium followed
Level 0	Intact Coating	No reaction
Level 1	Holiday/ Fe	Fe ²⁺ /Fe
Level 2	Holiday/Fe	Fe/ Fe ₃ O ₄
Level 3	Holiday/Fe	Fe ₃ O ₄ /Fe ₂ O ₃
Level 4	Holiday/Fe	Fe ²⁺ / Fe ₃ O ₄

During filed conditions we have the collection of indirect data and survey results, previously we aligned the information by using the location of each sense magnitude. The close interval survey indicates the E on and E off potentials. Different levels of severity were assumed to be classified with thermodynamic approach. The critical assumption is the E off conditions were able to capture the Cell potential and therefore the metastable surface condition could be assuming to follow the thermodynamic conditions at the interface.

A new classification could be establish based on the calculation and we can validate with the experimental results from the laboratory conditions.

Table 2 presents a new column that is severity due to thermodynamic condition originated from the Eoff potential. This new column will be correlated with the experimental testing, the experimental testing will extract the conditions based on the surface and the potential response. Also, there are functions that will be used to distinguish the severity from the indirect and survey methods.

Start Station	End Station	Ortal wall thickness, μ	Potential Redox (mV)	Eh	pH	CO3 (mmol L-1)	HCO3 (mmol L-1)	Cl (mmol L-1)	SO4 (mmol L-1)	ON	OFF/Ecell	E Protection -850mV	Severity
0	725.39	0.252	-362	-162	5.454	0.516	4.902	8.25	0.090096	-0.6641	-0.5819	Lower	Level 2
0	725.39	0.252	-362	-162	5.454	0.516	4.902	8.25	0.090096	-0.6641	-0.5819	Lower	Level 2
0	725.39	0.252	-362	-162	5.454	0.516	4.902	8.25	0.090096	-0.6641	-0.5819	Lower	Level 2
0.01	725.39	0.252	-362	-162	5.454	0.516	4.902	8.25	0.090096	-0.6641	-0.5819	Lower	Level 2
0.01	725.39	0.252	-362	-162	5.454	0.516	4.902	8.25	0.090096	-0.6641	-0.5819	Lower	Level 2
0.56	725.39	0.252	-362	-162	5.454	0.516	4.902	8.25	0.090096	-0.6641	-0.5819	Lower	Level 2
0.66	725.39	0.252	-362	-162	5.454	0.516	4.902	8.25	0.090096	-0.6641	-0.5819	Lower	Level 2
0.66	725.39	0.252	-362	-162	5.454	0.516	4.902	8.25	0.090096	-0.6641	-0.5819	Lower	Level 2
0.9	725.39	0.252	-362	-162	5.454	0.516	4.902	8.25	0.090096	-0.6641	-0.5819	Lower	Level 2
1.4	725.39	0.252	-362	-162	5.454	0.516	4.902	8.25	0.090096	-0.6714	-0.5909	Lower	Level 2
2.03	725.39	0.252	-362	-162	5.454	0.516	4.902	8.25	0.090096	-0.6356	-0.5556	Lower	Level 2

Table 2. Master file aligned with ROW properties, ILI results, indirect results, survey parameters and severity column condition.

Task 3: Bayesian machine learning to bridge the gaps in uncertainty quantification.

In the last quarterly report, the pipeline along the right-of-way has been divided into segments with unified length (100 meter) and each pipeline segment was linked with the nearby soil survey data, large scale vegetation and precipitation data, indirect inspection data (DCVG and CIPS), and in-line inspection data. After a thorough investigation of feature selection and dimension reduction, clustering analysis has been performed to extract homogeneous pipeline sections with similar soil environments. Once the entire pipeline has been segmented into regions of similar soil corrosivity, for each cluster, the next step is to develop a classifier to determine the defect types using pipeline indirect inspection data so that it can be applied to real-world practices in a cost-effective manner. In this project, existing data from Close Interval Potential Survey (CIPS) and Direct Current Voltage Gradient (DCVG) along with in-line inspection is used for supervised classification. The former two datasets are predictors and the severity assessments from in-line inspection are considered as responses.

Supervised learning is a machine learning process where an algorithm is implemented to train a model in terms of finding a mapping that connects training data input and output. In the current context, the indirect inspection data is used to classify defects. DCVG aims at locating defects in the pipeline coating. Voltage gradients are measured by a single operator using two reference electrodes in contact with the soil at a constant distance. While this technique is accurate in locating defect, it lacks the capability of predicting the defect severity. The results from DCVG along with ILI is used to determine abnormal points along the right of way. CIPS is a complimentary technique used along with DCVG to evaluate the criteria for adequate cathodic

protection (CP), and identify local deficiencies in the system. As per requirement NACE standard recommendations a value more negative than -850 mV while less negative than $-1.2V$ pipe-to-soil potential would be adequate to prevent significant corrosion. Regions with polarized pipe-to-soil potentials more negative than $-1.2V$ are regions of cathodic overprotection. Over protection causes acceleration of coating degradation and the possibility of hydrogen induced cracking. Figure 4 shows plot of CIPS on-off potential in each cluster with the level of CP protection along with metal loss points and Figure 5 gives the distribution of metal loss in each CP protected regions. From the figure we see that there is metal loss in each of the three regions and Figure 5 shows that in most clusters there is more metal loss in regions with CP. This proves that there are other factors that hinders CP and aids corrosion.

Scatter plot of CIS ON-OFF data per cluster

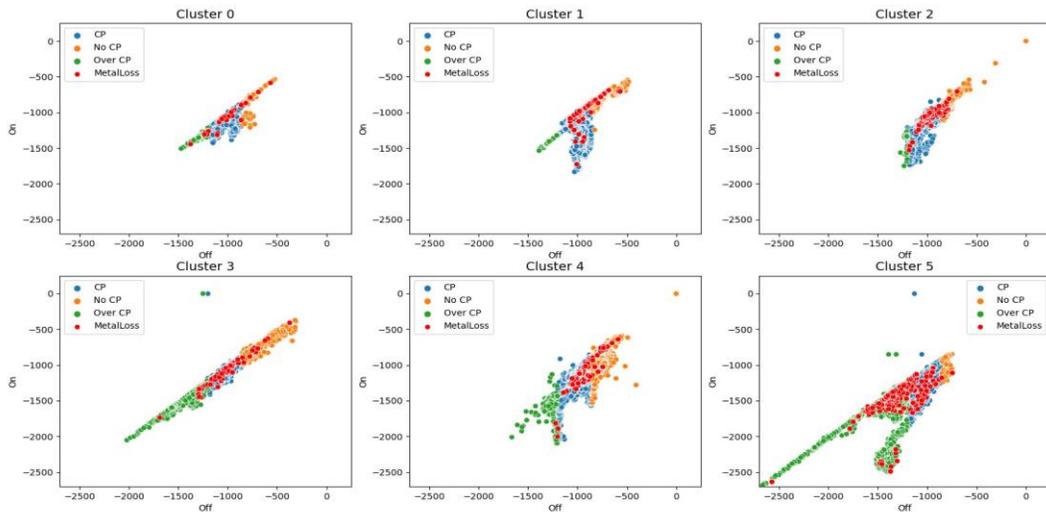


Figure 4: CIPS on-off potential per cluster

Metal loss in each CP region per cluster

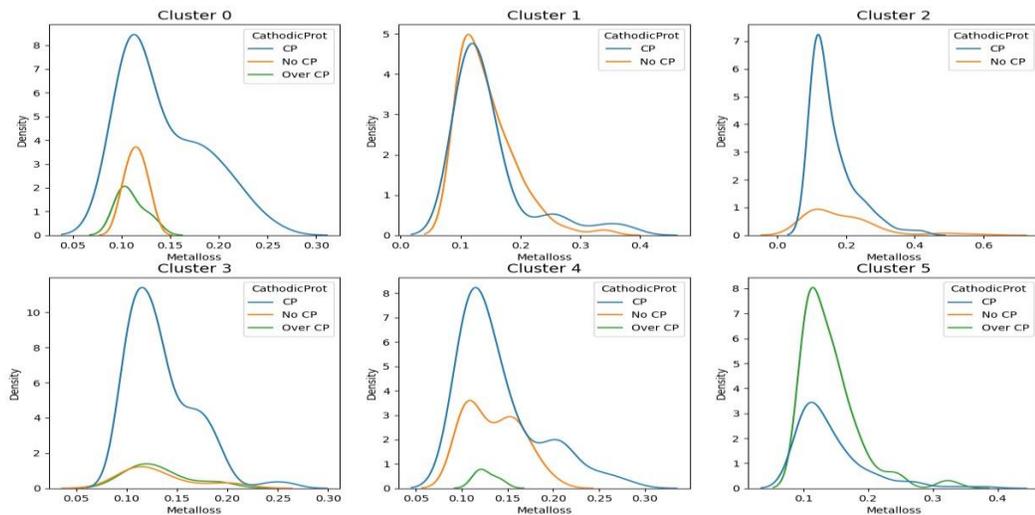


Figure 5: Density distribution of metal loss in each CP regions per cluster

It has been shown that the application of CP polarizes steel in the electronegative direction, while raising the pH at the interface at the same time leading to the formation of passive layer. CIPS measurement of potential more positive than CP protection criteria in these regions does not necessarily indicate corrosion. In unaerated environments, the steel polarized potential is directly dependent on the interfacial pH produced by CP. Hence, pH level is an important feature in detecting defects with CP criteria. Figure 6 shows the metal loss with pH severity. According to literature severity 3 corresponds to a pH between 6.0-8.0 and 4 corresponds to pH less than 6.0.

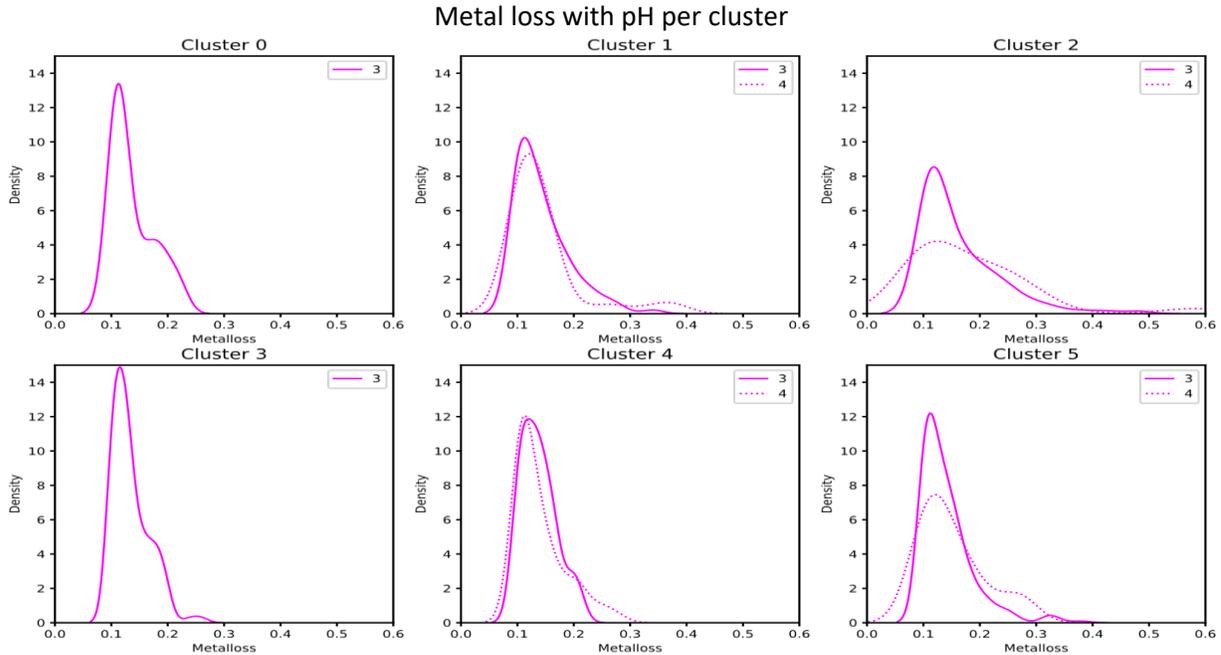


Figure 6: Density distribution of metal loss according to pH severity per cluster, severity 3 corresponds to a pH between 6.0-8.0 and 4 corresponds to pH less than 6.0.

Similarly, as soil resistivity increases the on potential becomes more positive indicating CP protection of pipeline has decreased. This is due to reduction in the amount of current that can reach the pipeline at higher soil resistivity. The difference in on off potential yields a dip at locations of coating defects. This dip becomes smaller as soil resistivity increases. Hence high soil resistivities may hide the presence of a coating flaw. A poorly designed CP may cause structural damage and stray current is an important indicator. Stray currents are currents flowing in the soil from external sources and these are also influenced by soil resistivity. Failure between coating and steel pipe is pipeline disbondment. This leads to the formation of crevice and corrosion. Cl^- concentration and acidification are observed at the bottom of the disbonded region. pH at bottom of the disbonded coating reduced gradually. Cl^- are usually prone to diffuse into the bottom of corrosion pits or crevice or crack tip with a similar mechanism to pitting or crevice corrosion. Based on literature the CIPS measurements, along with soil pH, Cl^- , resistivity and pipe resistance is used as input features for supervised classification of abnormal data points into coating defect and metal loss defects.

Some of the most widely used classification techniques are K-nearest neighbor (KNN), random forest, Naive Bayes and support vector machines (SVM). A major problem in data mining of real-world problems is class imbalance. Imbalanced data is where there is a significant difference between the class prior rates, the probability a particular data belongs to a particular class. The

most prevalent class is called the majority class, while the rarest class is called the minority class. For traditional classification models classifying imbalance data is a great challenge and often provide sub optimal classification results. The main difficulties with class imbalance are,

- Small sample size: The number of minority class samples are insufficient to train the classifier resulting in poor generalization and possible data overfitting.
- Small disjuncts: In some cases, the minority class may be represented by a number of sub concept and may affect classification performance.
- Class overlapping: Class overlapping between minority and majority classes deteriorates the total classification accuracy. As the minority class is under-represented in the data set, it is more likely underrepresented in the overlapping region.

Literature shows that random forest classifier performs well in dealing with datasets having large class imbalance. A random forest is a classifier consisting of a collection of trees structured classifiers $\{h(\mathbf{x}, \Theta_k), k = 1, \dots\}$ where the $\{\Theta_k\}$ are independently and identically distributed random vectors and each tree casts a unit vote for the most popular class at input \mathbf{x} . These tree voting procedures are collectively defined as random forests. There are two basic strategies for addressing imbalanced data that are data preprocessing and cost-sensitive learning. In preprocessing a resampling approach in which the training data are modified to produce a balanced data distribution that allow classifiers to perform in a similar standard. Cost-sensitive learning can be incorporated both at the data level and at the algorithmic level by assuming higher costs for the misclassification of minority class samples with respect to majority class samples. Adopting a cost sensitive learning by assigning a weight to each class, with the minority class given larger weight the random forest classifier can be made to perform better for imbalanced dataset. Hence, in this project, random forest classifier was used to classify defects as metal loss and coating defects. More details regarding the model validations are shown in the report of Task 4 below.

Task 4: Finalize and evaluate/validate the model.

One appropriate metric that could be used to measure the performance of classification over imbalanced datasets is the Receiver Operating Characteristic (ROC) curve. In this curve the tradeoff between the benefits (True Positive rate) and costs (False Positive rate) can be visualized and acknowledges the fact that the capacity of any classifier cannot increase the number of true positives without also increasing the false positives. The Area Under the ROC Curve (AUC) corresponds to the probability of correctly identifying which of the two stimuli is noise and which is signal plus noise. The ROC-AUC score is the area under the ROC curve. The precision score explains what proportion of positive identifications was actually correct. The measures of the quality of classification can be further analyzed using confusion matrix, which records correctly and incorrectly recognized examples for each class. Hence precision, ROC-AUC score and confusion matrix are used as evaluation matrices for classifier performance. The entire abnormal dataset was divided into training and testing sets. The training sets was used to do a cross validation of the classification model and the testing dataset was used to do a blind test of the model. In order to compare the performance other than random forest classifier, support vector machine and relevance vector machine were tested on the dataset. Figure 7, 8 shows the performance of the three classifiers using cross validation. Figure 9 shows the confusion matrix of the random forest classifier in cross validation. Figures 10, 11 and 12 are the results of blind validation on testing dataset. Comparing all results, we see random forest gives a stable performance in all cluster groups compared to SVM and RVM. The results of classification will

be further investigated using the results from laboratory tests, which will give us a deeper understanding of other underlying factors.

Cross validation results:Precision

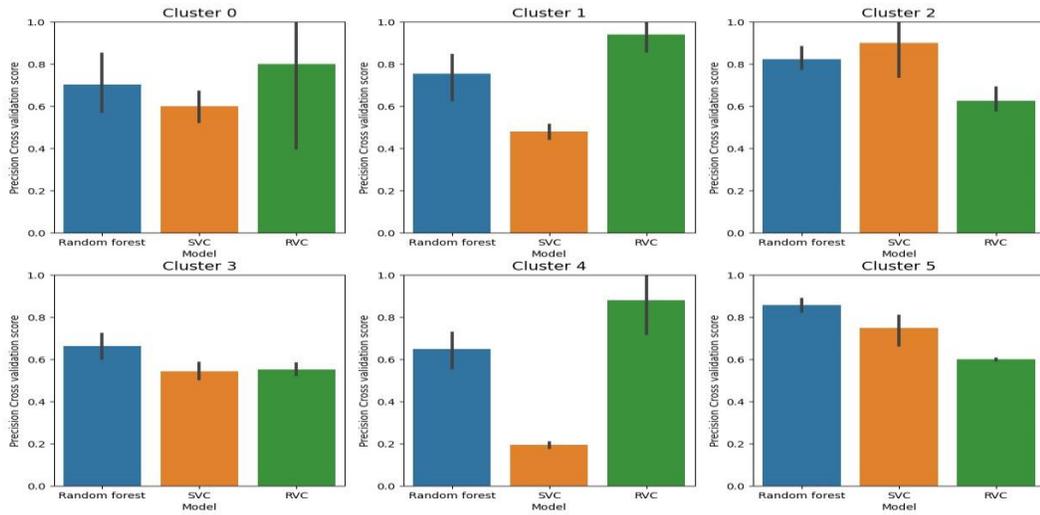


Figure 7: Cross Validation results: Precision

Cross validation results:ROC_AUC

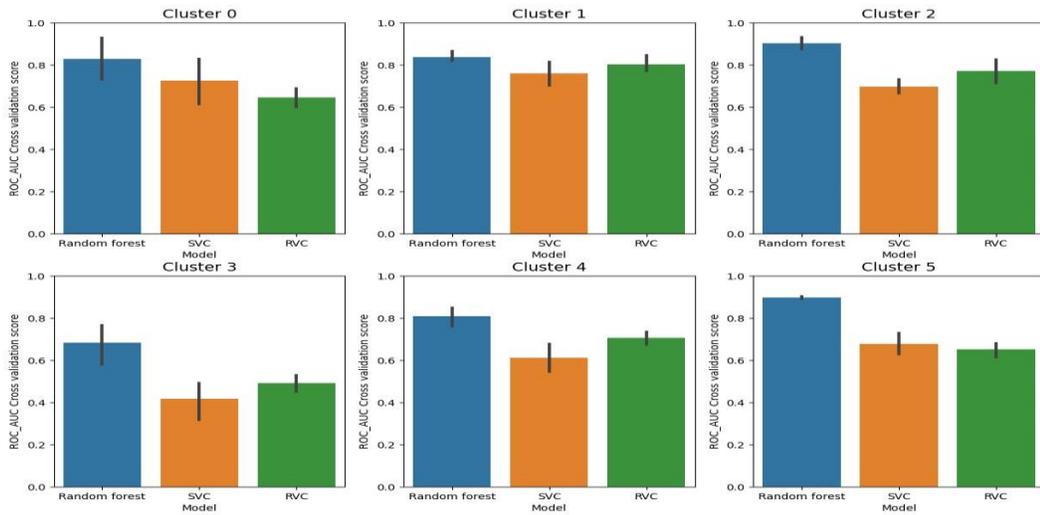


Figure 8: Cross Validation results: ROC-AOC score

Confusion matrix:Random forest

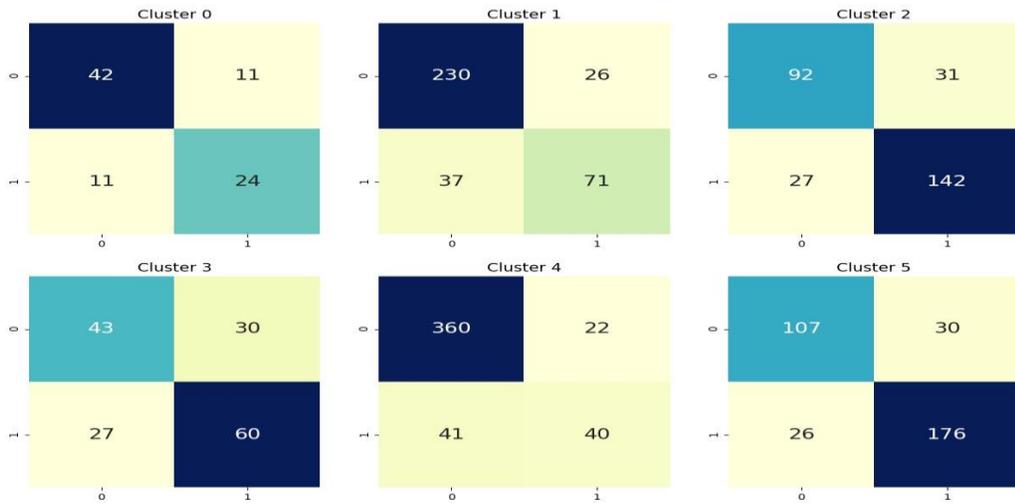


Figure 9: Random Forest Confusion Matrix

Validation results:Precision

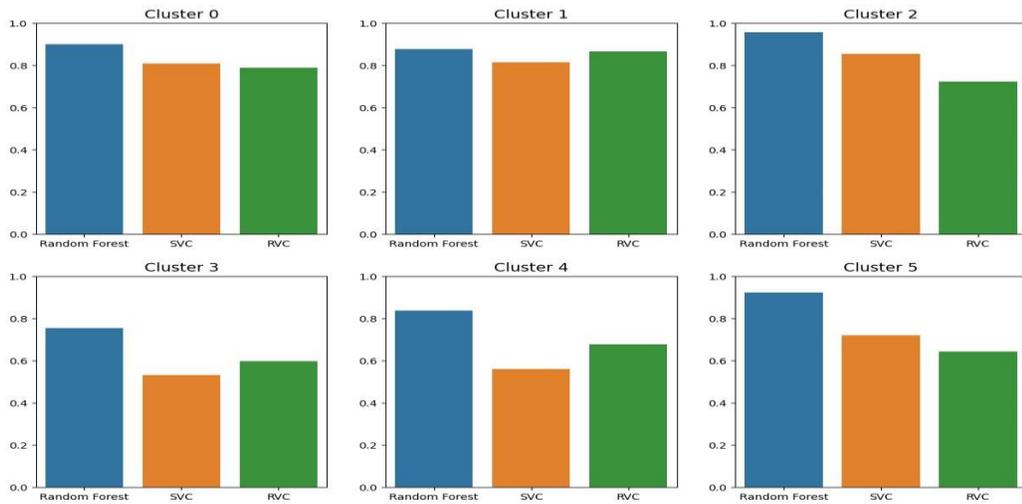


Figure 10: Validation results: Precision

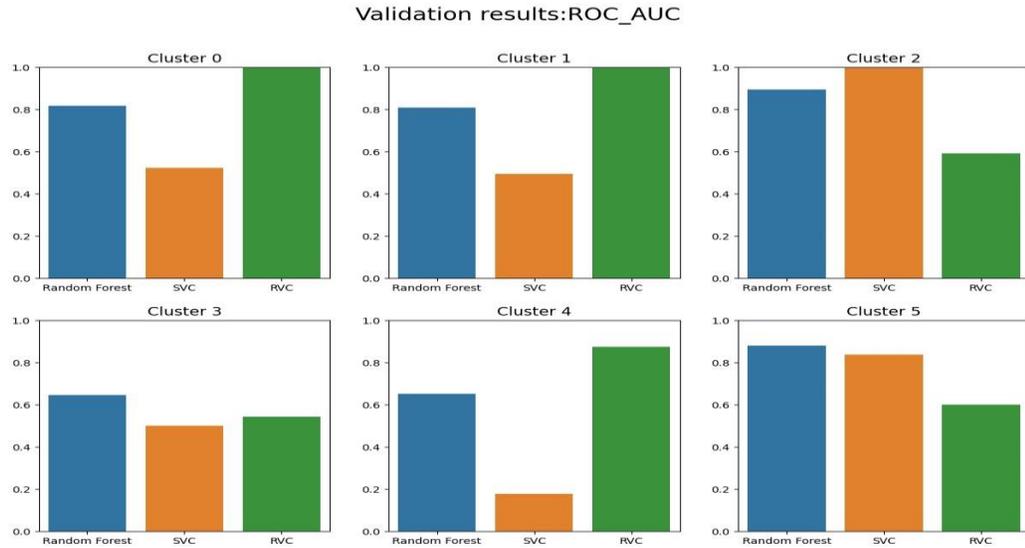


Figure 11: Validation results: ROC-AOC

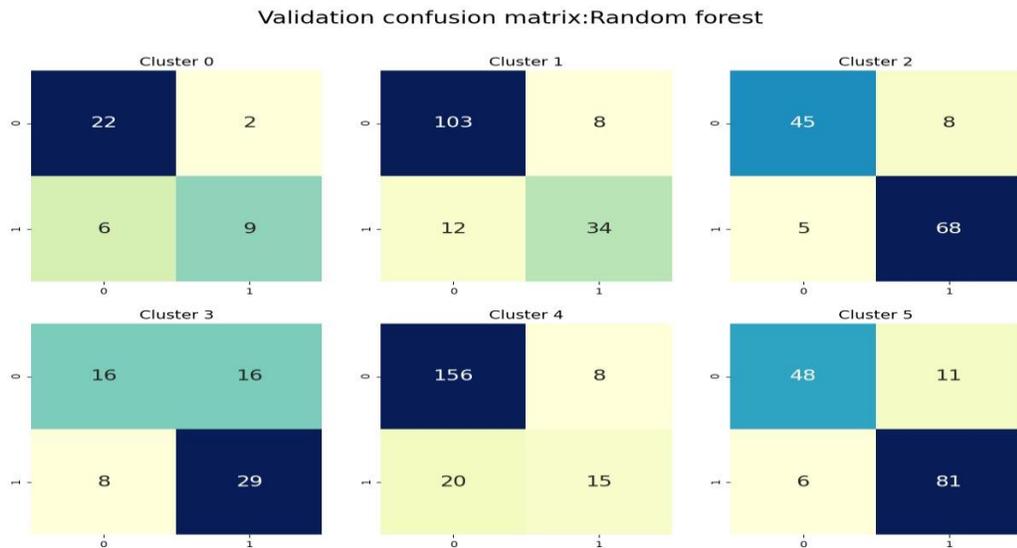


Figure 12: Validation results: Confusion Matrix

5: Project Schedule –

The project is on-schedule as originally-proposed.

During the following quarter, the team will perform the analysis of the experimental results, the correlations between field measurements and parameters founded in laboratory conditions to extract the information required for the machine learning method.

6. Publication

On December 31st, 2020, the peer review paper entitled: **Global and Local parameters characterizing and modeling External Corrosion for Steel Underground Pipelines: A**

review of Critical Factors was submitted to the Journal of pipeline science and engineering.

On December 16th, 2020, the final version of the peer reviewed conference paper entitled: **Mapping Indication Severity Using Bayesian Machine Learning from Indirect Inspection Data By Considering The Impact Of Soil Corrosivity** was submitted to the NACE CORROSION 2021 waiting for the final approval.

Observations: The experimental analysis for task 2 still is underway due to a delayed influenced by the COVID-19 circumstances.

In the following month, we will combine the new findings from Task 2 with the supervised classification model developed in Task 3. To be more specific, the E-pH diagram and the semi-empirical models developed from Task 2 will be used as the governing law and integrated into the supervised learning model and the performance will be compared with that from the random forest.

We also will extract the laboratory information to build the tables that can be used for the machine learning methodology.