CAAP Annual Report (October 2020-December 2020)

Date of Report:	January 13, 2021
Prepared for:	U.S. DOT Pipeline and Hazardous Materials Safety Administration
Contract Number:	Cooperative Agreement #693JK31850011CAAP
Project Title:	Development of a prediction model for pipeline failure probability based on learning from past incidents and pipeline specific data using artificial neural network (ANN)
Prepared by:	Texas A&M Engineering Experiment Station
Contact Information:	Ben Patterson
For quarterly period ending:	December 31, 2020

Business and Activity Section

(a) Contract Activity

The contract activities have been continued.

(b) Status Update of Past Quarter Activities

Dr. Noor Quddus (Assistant Research Engineer) has been leading the team for the project. Graduate students working in the project are Mengxi Yu (5th year PhD candidate), Pallavi Kumari (4th year Ph.D. candidate), and Guanyang Liu (4th year Ph.D. candidate). A senior undergraduate researcher, Mason Boyd, has also been working. Syeda Zohra Halim, a postdoctoral research associate/assistant research engineer also contributed to some work.

(c) Cost share activity

We have met $19,842.88 in cost share thus far on this project. We have met another $25,160 in cost share awaiting the approval of the revised cost share budget.

(d) **Task 1:** *Development of methodology for creating root cause analysis reports*
Task 2: Selection of training samples and development of the learning algorithm
Both Task-1 and Task-2 have been continued.

1. Background and Objectives

Pipeline incidents data obtained from the PHMSA incident database and incident investigation reports are being analyzed to develop a predictive model based on artificial neural network. Since incident reports collect only one apparent cause and one sub-cause per incident, the interdependency or relationship among the causal factors cannot be identified. Incident descriptions and incident investigation reports are analyzed to develop a taxonomy and relationships among the causal factors. Then along with other data that contributed to past incidents

will be used to develop the artificial neural network (ANN) model as a predictive model. The overall objective is to develop a knowledge based predictive model to assess pipeline failure through:

I. Learning about causes behind pipeline failure: Conducting root cause analysis of past incidents to identify those factors that have to potential to contribute to failure. The findings are to be specific to the extent that they can be applied into a predictive model.

II. Implementation of learning to predict failure: Utilizing the learnings about contributing factors behind pipeline failure to develop a predictive model based on artificial neural networks that monitors current existing conditions to determine dynamic failure probability of a pipeline.

2. Analysis in the last Quarter

This quarter, the focus was to continue the development of natural language processing techniques analyzing the incident description and incident investigation reports (Task – I), and development of an artificial neural network model to predict pipeline failure (Task – II).

Development of natural language processing (NLP) model: Two different datasets have been used to develop the NLP model. About 3,600 incident descriptions of HL incident from 2010-2019 were used for the first NLP model. Each incident description is short with an average of 140 words. The dataset was first pre-processed with standard practices. Later domain knowledge was used to filter a few words to increase the effectiveness. K-means clustering technique and Co-occurrence network were applied on the dataset. Similar approach has applied to different sections of incident investigation reports (about 70) available at PHMSA websites. Most relevant sections are considered for preprocessing. In addition to K-means clustering technique and co-occurrence network, topic modeling algorithm is being used to extract useful information from the incident investigation reports. Both unsupervised and supervised approaches are considered.

About 40 incidents that are common to both datasets have been identified and compared. They are manually studied to understand how they are affecting the NLP results. This analysis will help us understand better the limitation of existing incident database and how to improve both the reporting system and investigation results.

Development of the artificial neural network (ANN) model: Two different approaches are being parallelly investigated. In the first approach, an ANN model has been developed with 737 corrosion incident data and about 100 relevant data fields were considered. The corrosion model has been targeted because there are a lot of corrosion model to compare and validate. All general information regarding the pipeline and corrosion specific data are used for this model development. The data fields are merged as much as possible without compromising the integrity of the model. Release type, release rate and cost of damage are considered as the output. Two-third of the data was used to train the model and one third of data was used to validate the model. Following this principle, another ANN model for equipment failure has been developed with 1500 equipment failure data. Three other individual models for external excavation, incorrect operation

and natural force damage are under development. In the second approach, an ANN model is being developed considering all failure causes. Same approach has been considered for input and output data and training and test data. The challenge is that failure data for all causes are not of the same order. We are exploring to use some external data sources for this approach.

Results

We presented the results at three conference papers virtually during these periods.

- Noor Quddus, Guanyang Liu, Mason Boyd, Analysis of Pipeline Incident Data and Investigation Reports Using Natural Language Processing (NLP), paper presented at Hazards30, 18-20 May, Manchester, UK
- Guanyang Liu, Mason Boyd, Pallavi Kumari, Syeda Zohra Halim, and Noor Quddus, An Intelligent Learning Framework for Analysis of Pipeline Incident Investigation Reports, Presented at Process Safety International Symposium 2020, College Station, October 2020.
- Pallavi Kumari and Noor Quddus, Causation Analysis of Pipeline Incidents Using Artificial Neural Network, Presented at MKOPSC Process Safety International Symposium, October 2020.

We submitted two more abstracts to 2021 Global Congress of Process Safety and PRCI Virtual Research Exchange 2021 conferences on the topic areas.

3. Future work

Future work will be continued on the following

- The NLP model development will be continued using incident investigation reports available at PHMSA and NTSB. The reports will be examined to identify the influence of causal factors, background factors and underlying factors to the pipeline failure and any interdependence among them.
- ANN model development will be continued including other failure causes and predict the future incident.
- Finally, the extracted information of NLP code will be included in the ANN model to improve the prediction.