Developing a predictive model using multivariate analysis and Long Short-Term Memory (LSTM) to assess corrosion degradation in mining pipeline thickness.

Kalidou Moussa Sow, Nadia Ghazzali
University of Quebec at Trois-Rivieres, Department of Mathematics and Computer Science
Kalidou.Moussa.Sow@uqtr.ca
Nadia.Ghazzali@uqtr.ca

Abstract

Pipeline corrosion has significant impacts on the human, economic, and natural environment. To help better detect and prevent it over time, in this paper, we propose a multivariate approach using machine learning. More precisely, we propose to study the evolution of the thickness of the mining pipeline using a multivariate approach and to implement a predictive model using the Long Short-Term Memory (LSTM) artificial neural network. Indeed, LSTM is a specific recurrent neural network (RNN) architecture designed to model temporal sequences. The proposed predictive model achieved an accuracy of 80% and a loss of 0.01 and was able to predict variations in eight thickness measurements over one hundred days.

Introduction

For many decades, pipelines have served as the most efficient and secure means of transporting materials globally. Any breakdown in pipeline transmission systems directly affects the economics of the material industry (El-Abbasy et al. 2014). Over the years, researchers have delved into studying models of these failures (Lam and Zhou 2016), (Valor et al. 2013). The majority of inquiries into assessing various failure modes in oil and gas pipelines indicate that corrosion stands out as one of the most prevalent causes of failures in transmission systems.

Pipeline corrosion leads to degradation and reduced pipeline thickness. A high-performance model is needed to quickly identify the risk of corrosion and effectively halt its spread, thus restoring the pipeline's original thickness.

(Hussein Farh et al. 2023) classified the cause of corrosion into three categories: 1) environmental factors; soil, external, and stray current factors; 2) pipe factors, and 3) operational factors. They used the fuzzy analytical hierarchy process to represent the impact of the three factors on pipeline corrosion. They found that operational factors had the highest relative weight (0.428), followed by environmental factors (0.337).

(Scully et al. 2008) explained that 60% of failures in Mexican oil and gas transportation systems are caused by pitting corrosion on the external walls of pipelines. This fact of high incidents occurring due to corrosion relates to the complexity of the environment that surrounds pipelines, including a large variety of soil properties, water, and transported products using the pipeline (Oil or Gas).

As a result of the advancements in machine learning (ML) and deep learning (DL), there has been significant interest in data-driven model-based detection methods for pipeline erosion-corrosion monitoring. The artificial neural network (ANN), particularly the backpropagation neural network (BPNN), is widely employed for predicting the erosion-corrosion rate in pipelines (Zhang, Du, and Cao 2015). (Tian, Gao, and Li 2006) adopted the ANN model to detect the corrosion of the submarine oil pipeline under four degrees (no, mild, moderate, serious) based on the original input data collected by the ultrasonic sensor and flux leakage sensor.

This paper aims to build a predictive model of the corrosion degradation of a pipeline used to convey water in the mining area of the Quebec metallurgy center. Work has already been done on the same data used in this paper. (Dia, Ghazzali, and Gambou Bosca 2022) have applied an unsupervised neural network, self-organizing maps (SOM), to study the impact of corrosion assessed by periodic ultrasonic inspections. They combined SOM and hierarchical clustering to detect the extent of corrosion in a mining pipeline. This paper builds upon the research conducted by (Dia, Ghazzali, and Gambou Bosca 2022) by introducing a novel approach that incorporates multivariate analysis and LSTM.

Related Work

Long Short-Term Memory is a specific recurrent neural network (RNN) architecture that was designed to model temporal sequences. It was developed in 1997 by Sepp Hochreiter and Jürgen Schmidhuber (Hochreiter and Schmidhuber 1997) to solve the vanishing gradient problem present in traditional RNNs. Its relative insensitivity to interval length is its advantage over other RNNs, hidden Markov models, and other sequence learning methods. LSTM is very good at predicting in a time series (Lara-Benítez et al. 2020). It could extract patterns from sequential data and store these patterns in internal state variables. Each LSTM cell can retain important information for a longer period when it is used. This information property allows the LSTM to perform well in classifying, processing, or predicting complex dynamic sequences (Mateus et al. 2021), which makes LSTM a very used model in the literature.

(Salman et al. 2018) applied an LSTM model to weather variable data collected by weather underground at Hang Nadim Indonesia. They added an intermediate variable signal on the memory block cell of LSTM and their model performs better than other LSTM models with an accuracy of 0.8060 and a Root Mean Square Error (RMSE) of 0.0775. (Nelson, Pereira, and De Oliveira 2017) used LSTM networks to predict future trends of stock prices based on the price history, alongside technical analysis indicators to an average of 55.9% accuracy when predicting if the price of a particular stock is going to go up or not shortly. (Di Persio and Honchar 2016) compares the performance of LSTM and Multilayer Perception (MLP) to their own proposed method based on a combination of wavelets and Convolutional Neural Networks (CNN), which outperforms both but has very close results to the LSTM network. (Karmiani et al. 2019) compared LSTM with Support Vector Machine (SVM), backpropagation, and Kalman filter for the stock market for different numbers of epochs varying from 10 to 100. LSTM was found to have high accuracy and low variance. (Chen, Zhou, and Dai 2015) used an LSTM model on the historical data of the Chinese stock market. They trained the LSTM model on 900000 sequences and tested it using the other 311361 sequences. Compared with the random prediction method, their LSTM model improved the accuracy of stock returns prediction from 14.3% to 27.2%. These efforts demonstrated the power of LSTM in predicting China's dynamic and highly unpredictable stock market.

LSTM has also proven itself in the analysis and prediction of corrosion pipelines Since the data collected in this area can be considered as time series. (Li et al. 2022) combined a new swarm intelligence optimization algorithm called SSA and a LSTM model to predict the maximum pitting corrosion depth of subsea oil pipelines. The comparison of their SSA-LSTM method with the LSTM alone shows that the new model SSA-LSTM performed superior in prediction accuracy and robustness. They used RMSE, Mean Absolute Error (MAE), Mean Squared Error (MSE) and Mean Absolute Percent Error (MAPE) as evaluation parameters to measure the performance of their model. The proposed hybrid model (SSA-LSTM) obtained the best performance with the smallest evaluation parameter values (RMSE = 0.0607 , MAE = 8.84% , MSE = 0.36%, MAPE = 9.58%).

Data processing

Data Collection

The data for this study were obtained from the Quebec Metallurgy Centre in collaboration with Agnico Eagle Mine Goldex. The range of data for this study was from year 2016 to year 2023. The database is composed of sixteen process variables which are collected every five minutes and eight pipe thickness variables which are measured using a probe installed in the pipe and recollected within 24 hours or more. Table 1 shows the distribution of process data as well as their mean and standard deviation (std). The number of registrations received for sixteen variables is 179989 and 635 for each thickness measurement. The nominal thickness of the pipeline is between 5.5mm and 6.5mm.

In contrast to (Dia, Ghazzali, and Gambou Bosca 2022), which aggregated the eight thicknesses by calculating their average, in this paper, the eight thicknesses and 16 process variables are considered to predict future measurements of the eight thicknesses.

Area	Parameter	Mean	Std
Alimentation	Tonnage Sag	337.88	61.74
Flotation sector	pulp flotation temperature	25.4	5.89
	pH flottation	9.08	0.26
Pipeline	residue flow	431.14	113.4
	% solid residue	24.98	15.7
	Calculated residual TPH	156.25	132.02
	Pressure Km 0	2095	737.3
	Temperature Km 0	18.89	6.7
	Pressure Km 14	430.36	446.66
	Temperature Km 14	18.09	19.87
Thompson River	flow rate m3/h	182.6	106.65
	Temperature	11.01	6.59
Sedimentary Basin	flow rate m3/h	70.27	15.99
	Temperature	12.78	7.55
South Park	flow rate m3/h	166.4	101.6
	Temperature	7.8	6.3

Tab. 1: The process data, their mean and their standard deviation

Data Analysis

Figure1 shows the distribution of the eight thicknesses. The thickness data varies between 4mm and 7mm, except for the boxplot for thickness 1, which shows thickness measurements greater than 40 and others equal to 0, indicating that there are input errors for thickness 1. To deal with these missing or outliers, we will replace them with a moving average to reduce noise and maintain the trend in thickness values.

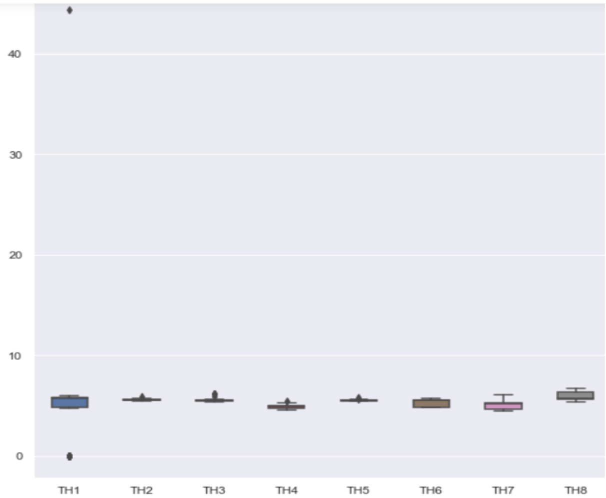


Fig. 1: Boxplot of the Eight thicknesses

Figure 2 shows the thickness data's evolution according to collection dates. Dates are represented in French (janv. January, avr. April, juil. July, and oct. October). After applying the various transformations, the values obtained vary between 4mm and 7mm.

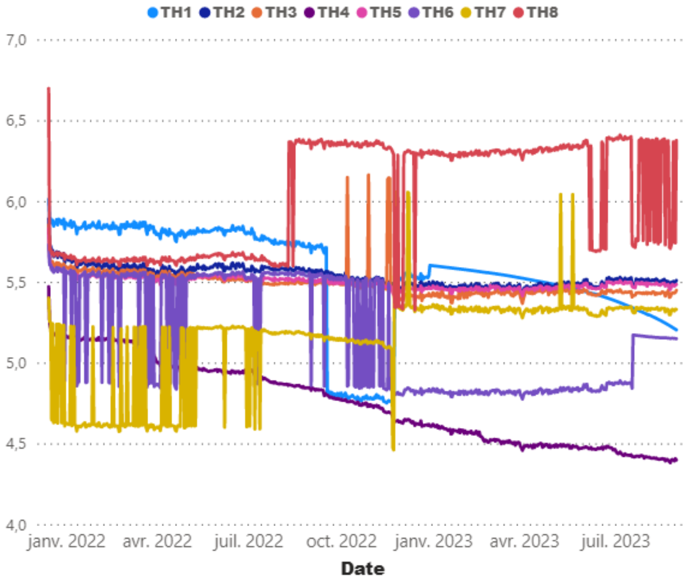
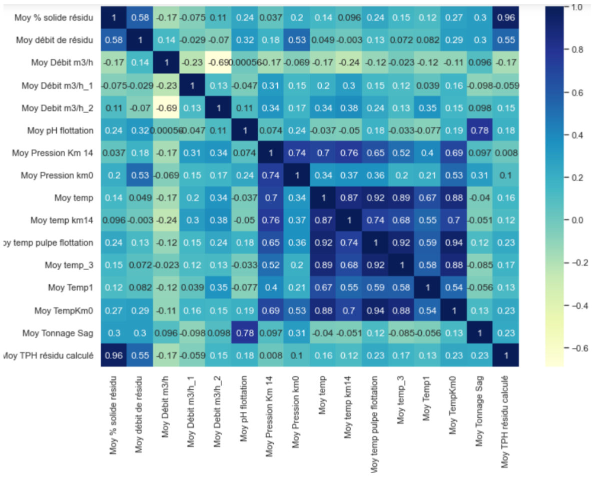


Fig. 2: The evolution of thickness measurements after the application of the transformations

The correlation matrix (see Figure 3) shows that some measurements are strongly correlated. It can be seen that the % solid of the residue and the Ton Per Hour (TPH) of the calculated residue are positively correlated with a Pearson coefficient of 0.96 and also the Ph of the flotation and the flow rate m3/h are negatively correlated. It is therefore interesting to reduce the data by performing a principal component analysis (PCA).



	Moy % solide résidu	Moy débit de résidu	Moy Débit m3/h	Moy Débit m3/h_1	Moy Debit m3/h_2	Moy pH flottation	Moy Pression Km 14	Moy Pression km0	Moy temp	Moy temp km14	Moy temp pulpe flottation	Moy temp_3	Moy Temp1	Moy TempKm0	Moy Tonnage Sag	Moy TPH résidu calculé
Moy % solide résidu	1	0.58	-0.17	-0.075	0.11	0.24	0.037	0.2	0.14	0.096	0.24	0.15	0.12	0.27	0.3	0.96
Moy débit de résidu	0.58	1	0.14	-0.029	-0.07	0.32	0.18	0.53	0.049	-0.003	0.13	0.072	0.082	0.29	0.3	0.55
Moy Débit m3/h	-0.17	0.14	1	-0.23	-0.69	0.00056	-0.17	-0.069	-0.17	-0.24	-0.12	-0.023	-0.12	-0.11	0.096	-0.17
Moy Débit m3/h_1	-0.075	-0.029	-0.23	1	0.13	-0.047	0.31	0.15	0.2	0.3	0.15	0.12	0.039	0.16	-0.098	-0.059
Moy Debit m3/h_2	0.11	-0.07	-0.69	0.13	1	0.11	0.34	0.17	0.34	0.38	0.24	0.13	0.35	0.15	0.098	0.15
Moy pH flottation	0.24	0.32	0.00056	-0.047	0.11	1	0.074	0.24	-0.037	-0.05	0.18	-0.033	-0.077	0.19	0.78	0.18
Moy Pression Km 14	0.037	0.18	-0.17	0.31	0.34	0.074	1	0.74	0.7	0.76	0.65	0.52	0.4	0.69	0.097	0.008
Moy Pression km0	0.2	0.53	-0.069	0.15	0.17	0.24	0.74	1	0.34	0.37	0.36	0.2	0.21	0.53	0.31	0.1
Moy temp	0.14	0.049	-0.17	0.2	0.34	-0.037	0.7	0.34	1	0.87	0.92	0.89	0.67	0.88	-0.04	0.16
Moy temp km14	0.096	-0.003	-0.24	0.3	0.38	-0.05	0.76	0.37	0.87	1	0.74	0.68	0.55	0.7	-0.051	0.12
oy temp pulpe flottation	0.24	0.13	-0.12	0.15	0.24	0.18	0.65	0.36	0.92	0.74	1	0.92	0.59	0.94	0.12	0.23
Moy temp_3	0.15	0.072	-0.023	0.12	0.13	-0.033	0.52	0.2	0.89	0.68	0.92	1	0.58	0.88	-0.085	0.17
Moy Temp1	0.12	0.082	-0.12	0.039	0.35	-0.077	0.4	0.21	0.67	0.55	0.59	0.58	1	0.54	-0.056	0.13
Moy TempKm0	0.27	0.29	-0.11	0.16	0.15	0.19	0.69	0.53	0.88	0.7	0.94	0.88	0.54	1	0.13	0.23
Moy Tonnage Sag	0.3	0.3	0.096	-0.098	0.098	0.78	0.097	0.31	-0.04	-0.051	0.12	-0.085	-0.056	0.13	1	0.23
Moy TPH résidu calculé	0.96	0.55	-0.17	-0.059	0.15	0.18	0.008	0.1	0.16	0.12	0.23	0.17	0.13	0.23	0.23	1

Fig. 3: Correlation matrix

After applying the PCA, analysis of the percentage of variation explained shows that from the ninth dimension onwards, we've already reached more than 99% of the variance explained (see Figure 4) so we'll keep only nine dimensions for the rest of the analysis.

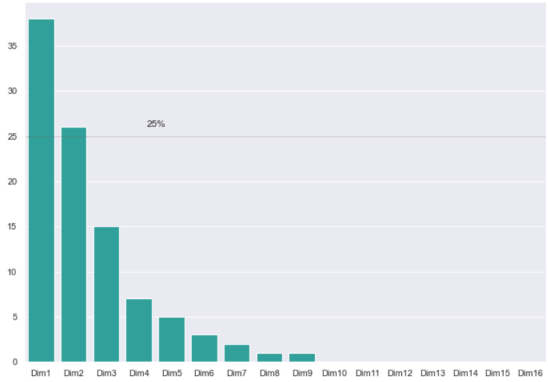


Fig. 4: percentage of variation explained by dimension

Long short-term memory

Long Short-Term Memory (LSTM) Networks allow to learn long-term dependencies. They are explicitly designed to avoid the long-term dependency problem.

An LSTM network has three gates that update and control the states of the cells: the Forget gate, the Input gate, and the Output gate.

The Forget gate (see Figure 5) is responsible for deciding to let information pass. State 0 corresponds to ”keep complete information” while state 1 represents ”Totally get rid of the information”. It is defined by the following equation:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{1}$$

where:

- W_f represents the weight matrix associated with the Forget gate.

- $[h_{t-1}, x_t]$ designates the concatenation of the current entry and the previous hidden state.
- b_f is the bias.
- σ is the sigmoid function.

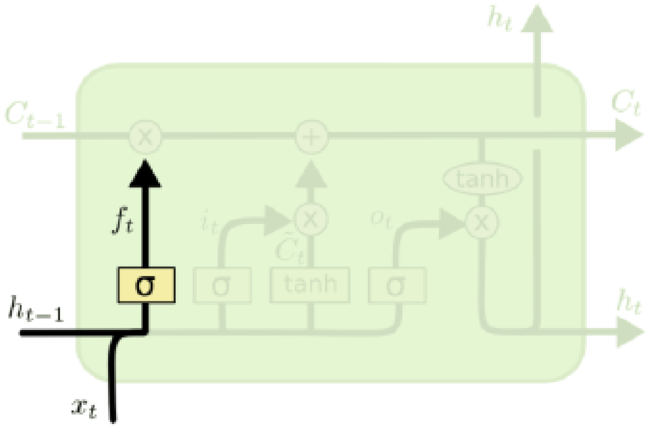


Fig. 5: Forget gate (Oinkina and Hakyll 2015)

The Input gate (see Figure 6) controls what new information will be encoded in the cell state, given the new input information. The information is regulated using the sigmoid function and filters the values to be retained in the same way as the forgetting gate, using the inputs h_{t-1} and x_t. Next, a vector is created using the tanh function, which gives an output from -1 to +1, containing all possible values of h t-1 and x_t. Finally, the vector values and the regulated values are multiplied to obtain useful information.

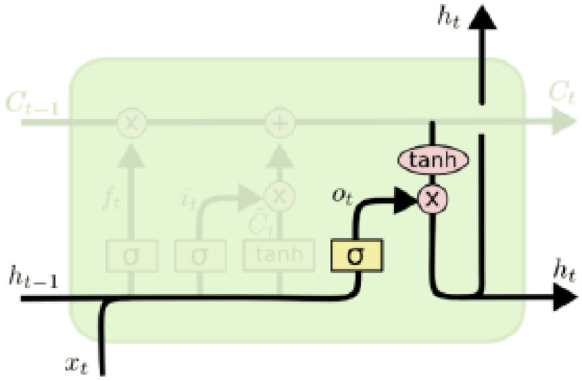


Fig. 6: Input gate (Oinkina and Hakyll 2015)

The input gate equation using the hyperbolic tangent function (tanh) is as follows:

$$\tanh(W_c[h_{t-1}, x_t] + b_c) \otimes \sigma(W_i[h_{t-1}, x_t] + b_i) \qquad (2)$$

where:

- W_c represents the weight matrix associated with the Input gate.
- $[h_{t-1}, x_t]$ designates the concatenation of the current entry and the previous hidden state.
- $\otimes$ is the element-by-element product.
- σ is the sigmoid function.

The Output gate (see Figure 7) controls which information encoded in the cell state is sent to the input network at the next time step, this is done via the output vector h(t). First, a vector is generated by applying the tanh function to the cell. Next, the information is regulated using the sigmoid function and filtered by the values to be retained using the inputs h_{t-1} and x_t. Finally, the vector values and the regulated values are multiplied and sent as output and input to the next cell. The output gate equation is as follows:

$$h_t = o_t \otimes \tanh(c_t) \text{ where } o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (3)$$

where:

- W_o represents the weight matrix associated with the Output gate.
- $[h_{t-1}, x_t]$ designates the concatenation of the current entry and the previous hidden state.
- b_o is the bias associated with the Output gate.
- $\otimes$ is the element-by-element product.
- σ is the sigmoid function.

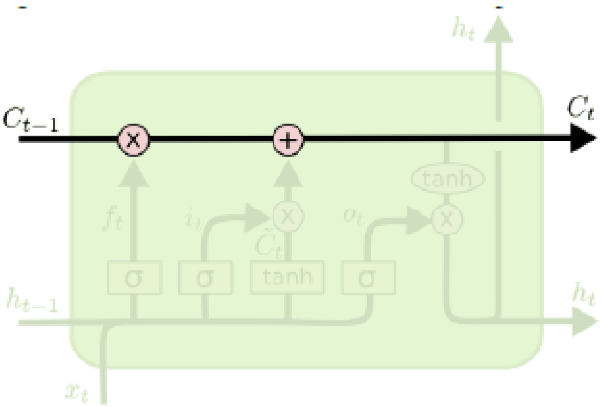


Fig. 7: Output gate (Oinkina and Hakyll 2015)

Results and Discuss

Model evaluation

Figure 8 shows the evolution of the loss function of the LSTM model over time. The decrease in the loss function proves that the LSTM model used has minimized the prediction errors during training. The performance of the model can also be measured by its accuracy.

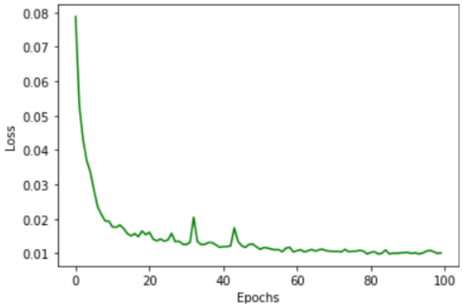


Fig. 8: Training Loss

During training, the model used achieved an accuracy score of 80%, which means that 80% of the values predicted by the model are correct (Figure 9).

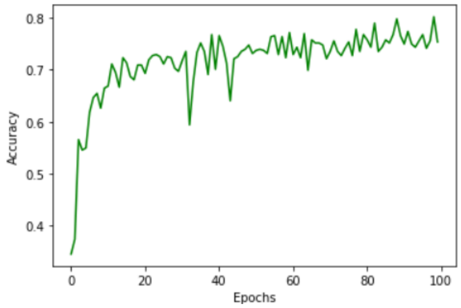


Fig. 9: Training Accuracy

Table 2 shows that using the PCA on the dataset increases the model's performance, with accuracy rising from 76% to 80 % and Loss decreasing to 0.100. This is because the PCA eliminates redundancies between variables.

Model	Accuracy(%)	Loss
LSTM	76	0.132
PCA + LSTM	80	0.100

Tab. 2: Comparison of LSTM and PCA+LSTM performance

Model predictions

Once the model has been trained, we will test it on the test data. To do this, we will predict the evolution of the thicknesses over the next 100 days. Figure 10 shows the evolution of the pipeline thickness 1, 2, 3, 5, and 6 predicted by the LSTM model for the next hundred days.

The thickness values of pipelines 1 and 6 are below the nominal thickness value (between 5.5 and 6.5) before the 40 days which means that the corresponding pipeline is affected by corrosion and should be monitored. And after 40 days, there is an increase in the thickness measurements of pipelines 1 and 6 which means that both pipelines (1 and 6) should be treated. The evolution curve of the thickness measurements predicted by the LSTM model of the pipelines (2, 3, and 5) shows an almost constant variation between 5.4mm and 5.6mm, which is just over the minimum thickness of 5.5mm. This suggests that regular inspections should be carried out over the next 100 days.

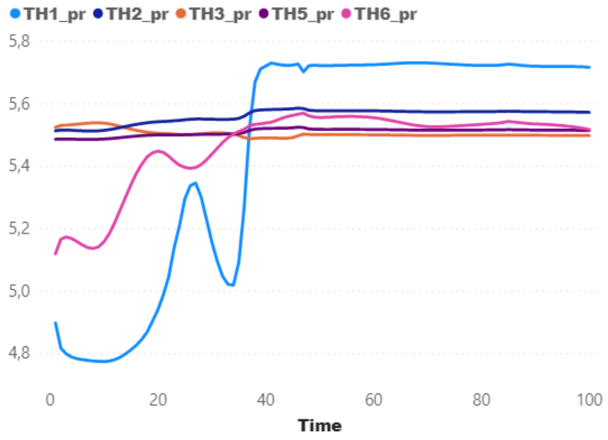


Fig. 10: Pipe thickness 1, 2, 3, 5 and 6 predicted by the model

The thickness measurements predicted by the model for pipelines 4 and 7 are above the nominal thickness (5mm), which means that pipelines 4 and 7 are already corroded, since the LSTM is monitoring the evolution of past measurements, the corresponding pipelines will have to be replaced or treated. (see Figure11)

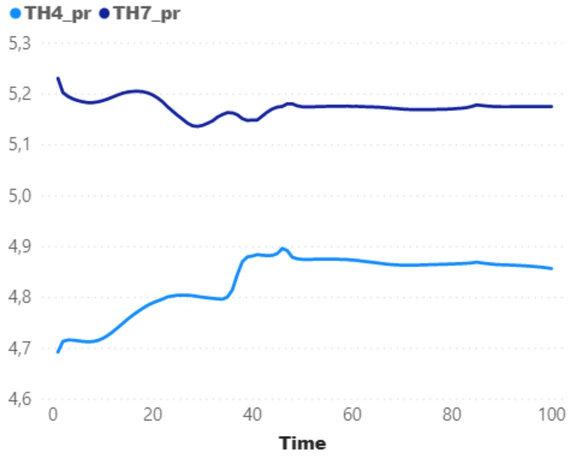


Fig. 11: Pipeline thickness 4 and 7 predicted by the model

The evolution of the thickness 8 curve predicted by the model shows a rapid fall in thickness 8 up to the 40th day, which means that there is a risk of loss of thickness due to the corresponding corrosion of the pipe. Even if the measurements of the thickness of the corresponding pipeline are between the nominal thicknesses, it must be treated before the 40 days to avoid the increase in corrosion which will correspond to the increase in the measurement of the thickness of pipeline 8.

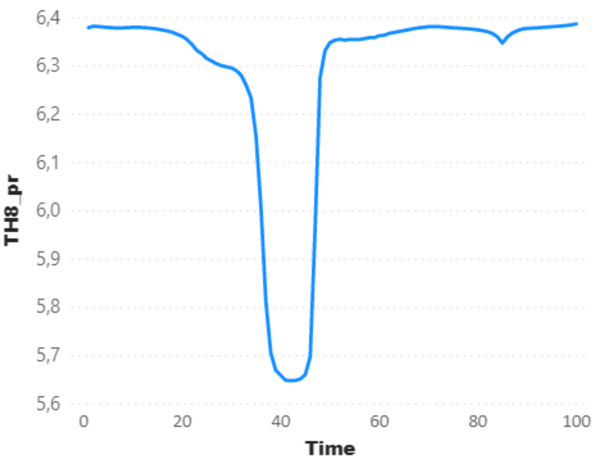


Fig. 12: Pipeline thickness 8 predicted by the model

Conclusion

Pipeline failure has attracted the attention of many research communities because of its significant impact on the global economy, leaks, explosions, and costly downtime. Many efforts to build corrosion prediction models have been proposed resulting in a vast number of publications available in the literature. However, the nature of corrosion pipeline is so complex that difficult to formulate in a single mathematical model.

In this paper, an LSTM model was used to predict the evolution and measurement of the thicknesses of an ore transport pipeline. The model developed was able to predict the evolution of different pipeline thickness values.

Future work will focus on the received data and combine other time series processing models with the LSTM.

Acknowledgement

We thank Agnico Eagle Goldex Mine for its support. This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Québec Fonds de recherche nature et technologies (FRQNT).

References

Chen, K.; Zhou, Y.; and Dai, F. 2015. A lstm-based method for stock returns prediction: A case study of china stock market. In *2015 IEEE international conference on big data (big data)*, 2823–2824. IEEE.

Di Persio, L., and Honchar, O. 2016. Artificial neural networks approach to the forecast of stock market price movements. *International Journal of Economics and Management Systems* 1.

Dia, A. K.; Ghazzali, N.; and Gambou Bosca, A. 2022. Unsupervised neural network for data-driven corrosion detection of a mining pipeline. In *The International FLAIRS Conference Proceedings*, volume 35.

El-Abbasy, M. S.; Senouci, A.; Zayed, T.; Mirahadi, F.; and Parvizsedghy, L. 2014. Artificial neural network models for predicting condition of offshore oil and gas pipelines. *Automation in Construction* 45:50–65.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Hussein Farh, H. M.; Ben Seghier, M. E. A.; Taiwo, R.; and Zayed, T. 2023. Analysis and ranking of corrosion causes for water pipelines: a critical review. *npj Clean Water* 6(1):65.

Karmiani, D.; Kazi, R.; Nambisan, A.; Shah, A.; and Kamble, V. 2019. Comparison of predictive algorithms: backpropagation, svm, lstm and kalman filter for stock market. In *2019 amity international conference on artificial intelligence (AICAI)*, 228–234. IEEE.

Lam, C., and Zhou, W. 2016. Statistical analyses of incidents on onshore gas transmission pipelines based on phmsa database. *International Journal of Pressure Vessels and Piping* 145:29–40.

Lara-Benítez, P.; Carranza-García, M.; Luna-Romera, J. M.; and Riquelme, J. C. 2020. Temporal convolutional networks applied to energy-related time series forecasting. *applied sciences* 10(7):2322.

Li, X.; Guo, M.; Zhang, R.; and Chen, G. 2022. A data-driven prediction model for maximum pitting corrosion depth of subsea oil pipelines using ssa-lstm approach. *Ocean Engineering* 261:112062.

Mateus, B. C.; Mendes, M.; Farinha, J. T.; Assis, R.; and Cardoso, A. M. 2021. Comparing lstm and gru models to predict the condition of a pulp paper press. *Energies* 14(21):6958.

Nelson, D. M.; Pereira, A. C.; and De Oliveira, R. A. 2017. Stock market's price movement prediction with lstm neural networks. In *2017 International joint conference on neural networks (IJCNN)*, 1419–1426. Ieee.

Oinkina, and Hakyll. 2015. Comprendre les réseaux lstm.

Salman, A. G.; Heryadi, Y.; Abdurahman, E.; and Suparta, W. 2018. Single layer & multi-layer long short-term memory (lstm) model with intermediate variables for weather forecasting. *Procedia Computer Science* 135:89–98.

Scully, J. R.; Budiansky, N. D.; Tiwary, Y.; Mikhailov, A. S.; and Hudson, J. L. 2008. An alternate explanation for the abrupt current increase at the pitting potential. *Corrosion Science* 50(2):316–324.

Tian, J.; Gao, M.; and Li, J. 2006. Corrosion detection system for oil pipelines based on multi-sensor data fusion by improved simulated annealing neural network. In *2006 International Conference on Communication Technology*, 1–5. IEEE.

Valor, A.; Caleyo, F.; Hallen, J. M.; and Velázquez, J. C. 2013. Reliability assessment of buried pipelines based on different corrosion rate models. *Corrosion Science* 66:78–87.

Zhang, L.; Du, Y.; and Cao, A. 2015. The design of natural gas pipeline inspection robot system. In *2015 IEEE International Conference on Information and Automation*, 843–846. IEEE.