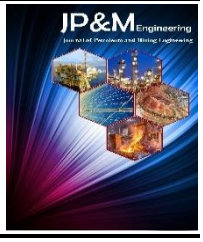
JP&M Engineering
Journal of Petroleum and Mining Engineering



# Comparative Analysis on the Prediction of Leak on Gas Pipeline Using Physical Models: Mathematical Models Versus Machine Learning Regression Models

Anthony O. Chikwe, Aniyom E. Ananiyom, Onyebuchi I. Nwanwe*, Jude E. Odo
Department of Petroleum Engineering, Federal University of Technology Owerri, Imo State, Nigeria
*Corresponding author e-mail: onyebuchi.nwanwe@futo.edu.ng

**Article Info**

Received 23 Sept. 2023
Revised 13 Mar. 2024
Accepted 17 May 2024

**Keywords**

Machine Learning, Leak Detection, Model, Gas pipeline, Mathematical Models, Prediction.

## Abstract

Pipeline leaks in the natural gas industry present multifaceted challenges, encompassing not only diminished product volume but also environmental degradation and potential catastrophic events such as explosions. Addressing these challenges requires a comprehensive approach, including the development and implementation of effective detection systems. Previous efforts have focused on physical surveys and the utilization of acoustic systems and pressure sensors to detect leaks promptly. However, recent advancements in technology have spurred interest in mathematical and machine learning models as potential solutions. This study delves into the comparative analysis of mathematical and machine learning models for leak prediction in gas pipelines, aiming to discern the most effective approach. Specifically, an existing mathematical model, derived from the Weymouth equation, is pitted against machine learning algorithms including random forest regressor, XGBoost, and voting regressor. Through rigorous evaluation, encompassing statistical error metrics, sensitivity analysis, and economic considerations, the study sheds light on the relative efficacy of these models. Ultimately, the findings not only contribute to enhancing leak detection capabilities but also underscore the transformative potential of machine learning in addressing complex industrial challenges.

## Introduction

The pipeline has been the major means of transportation for petroleum and its by-products from one place to another. It provides better efficiency in terms of fluid transportation compared to rail, trucks, and marine transportation. The transportation of fluids (Petroleum products) from the point of production to the end-users is now made possible and has led to the increase in the number of pipelines being constructed and laid [1). With this, it becomes very important for the mechanism flow in petroleum pipelines to be adequately studied and well understood.

Due to the toxic and hazardous nature of the products flowing through the pipeline, these products could cause accidents and environmental hazards if a leak occurs. These leaks are sometimes caused by complexities of environmentally or human-induced factors and disturbances generated along the pipeline flow network [2]. Due to the increasing awareness and empathy for the environment, most of the leakage from gas pipelines has shown cost-effectiveness, and the demand for reliable detection systems is very high. This implies that the financial costs usually incurred by the company are often significantly high, including the cleaning cost of the environment and the payment for pollution as stated by the Environmental Guidelines and Standards for the Petroleum Industry in Nigeria (EGASPIN), which was issued to the Department for Petroleum Resources (DPR) at the Ministry of Petroleum Products in 1991 [3].

Thus, the early detection of leakages in the gas pipeline offers several advantages amidst the economic advantages: safety of gas transportation, environment protection, gas quality protection, and avoiding pipeline breakages that could be used for subsequent transportation [4]. There are several methods for the detection of leaks in gas pipelines; software and sensor devices have been built for this purpose as well. Supervisory Control and Data Acquisition (SCADA) systems have been used severally for monitoring and control of numerous industrial and infrastructural processes in the oil and gas companies. It uses several computer-based algorithms during the transportation and distribution processes as applied in the oil and gas industries. It is useful in the detection of several anomalies to find unknown intrusions in the system [5].

Leakages in gas pipelines cause economic loss to companies due to damages caused to the pipelines and the pipe network, including maintenance purposes. Since gas has low density, leakage in gas pipelines could cause a huge risk to the inhabitants of the location and far beyond the point of the leak,

which can further result in an explosion. This study seeks to compare the use of physical models; a mathematical model modified from Weymouth horizontal gas pipeline equation and the use of intelligent models, which are machine learning regression models. The following metrics have been adopted for the evaluation of these models: statistical error metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Sensitivity analysis, economic analysis, etc., to validate the best models for the prediction or the detection of leaks in a gas pipeline. This study assumes a horizontal gas pipeline with no default upon installation.

## Definition of Terms

### Comparative analysis

Comparative analysis is typically defined as the comparison of two or more processes, documents, patterns, datasets, results, or other objects in order to select the best option that fits into the solution of a problem [6].

### Prediction

Prediction is referred to the output of an algorithm after it has been trained on historical dataset and applied to new data when forecasting the likelihood of a particular outcome [7]. Example is the prediction of where leak is likely to occur in a natural gas pipeline.

### Leakage

Leak as used in a pipeline is the situation where a system is designed to channel fluids from one place to another that is flawed in such a way that it losses some quantity of its contents before reaching its destination

### Natural gas

Natural gas is a naturally occurring mixture of gaseous hydrocarbons consisting primarily of methane in addition to various smaller amounts of other higher alkanes. Low levels trace of carbon dioxide, nitrogen, hydrogen sulfide and helium which are also present.

### Physical models

These are physical representation of the characteristics resemblance of a modeled system, with broad interest of examining the systems and studying the make-up of the system [8].

### Machine learning model

A machine learning model is a file which has been trained to recognize certain types of patterns [9]. A machine learning model is trained by providing a set of data with in-built algorithms for the learning of the patterns in the dataset [10].

### Regression models

Regression model provides a function that describes the relationship between one or more independent variables and a response, dependent, or target variable [11]. It could be linear, multiple, non-linear and stepwise regression models.

## Mathematical Models

A mathematical model is often described as a system by a set of variables and a set of equations used to establish relationship between variables [12]. It could be used to establish the relationship between independent variables and targets [13].

### Theories

Leak detection is important during transportation of natural gas through a pipeline, to help preserved the volume of natural gas being transported through that pipeline. It is expected that the transmission of gas is done without any leak experienced. but due to unforeseen circumstances, this ideal state can't be achieved 100%. Over the years, there have been several techniques to detect and control the leak effect during transmission gas pipeline.

a. The use of physical inspection: This involved the walking around the piping environment to detect by the sense and the use of trained animals to detect the leaks in those areas. But this method is limited in that it will not suffice for long distance transportation as it will be time consuming and stressful to the personnel involved

b. The Use of Sensors: This requires the installation of pressure sensors at intervals of the pipelines through which the gas pipeline is being transported. It helps acquire data from the pipelines for leak detection. This method alone alert where there is leak. But cannot be used to predict other leaks in the future

c. The use of Physics based models: Physics based models are models which build following several sets of physics and mathematical equations. These models are easy to compute and can be simulated using some software

d. Data driven models: These are data driven solutions with great impact as they can be used to predict future happenings as against other methods which are limited to time.

## Previous Work

Fluid flow along a natural gas pipeline make use of pressure and/or flow indicators at different sections of a pipeline, mostly only the extremes. During normal pipeline operation, there is usually steady state relationship among these indicators. Changes in these relationships will signal the occurrence of leaks. Volumetric balance is the most straightforward flow monitoring method. A leak alarm will be generated when the difference between upstream and downstream flow measurements changes by more than an established threshold. There exist several models developed for the purpose of detecting leak with conditions, assumptions and conclusions being established over time.

Considering the fact that the inlet flow rate measurements of a gas pipeline are not available and the conventional mass balance techniques cannot be used, Dinis and team gave a statistical method to the detection of leaks in subsea liquid pipelines. But his method has not been tested in gas pipelines. Dynamic model-based methods attempt to mathematically model the gas flow

within a pipeline. Using this model, flow parameters are calculated at different sections of the pipeline, and these parameters are measured as well. Then leaks can be detected by comparing the calculated and measured parameters by discretizing the pipeline model with non-uniform regions along the line. The limitation of their study is tied to the fact that the predictions are only done on seen or already existing pipelines which reduces the potential of the model to forecast on new pipelines [14].

Obibuike and his team developed a Mathematical model for the estimation of the accurate time of leak and the pressure at which leak occurs. This was done by modification of the basic equation of compressible fluid flow which include; continuity equations, momentum equations, energy equations, equations of state. They stated that leak occur when inlet mass is not equal to the outlet mass of the fluid flowing through the pipeline. After establishing this fact, they went ahead to develop a model that relates the time a leak occurs and the time it was detected. Thereby allowing for leak detection in their own view. Although the model performed with a high level of accuracy and an average error of 0.377% it can only handle limited observations as it may fail when introduced to a larger number of observations [2, 15-16].

In order to account for the limitations experienced during handling of larger number of datasets, Akinsete in 2019 proposed a simplified approach to detect and locate leaks using ANNs as an overall rating between the patterns of pressure and flow. They adopted ANN architectures of two levels [17]. The first level identifies the leakage, while the second level estimates precisely the magnitude and location of leaks. Artificial neural network was used to detect leaks of compressed air in a section of duct. The training of the neural model was performed using vibroacoustic signals picked up by a piezoelectric accelerometer. The optimization algorithm for training was the Levenberg-Marquardt, allowing a fast convergence of training for the ANN. From the results, they could detect 98% of cases of leakage and 99% in other situations with the generation of vibrations but no leak [17, 18].

Several other machine learning models [19] have been developed over time with which excellent performances, across classification through regression models [20]. This paper focuses on comparing the mathematical models to the machine learning models to evaluate the best performing models. The same dataset will be fed into the systems and predictions made on new dataset to evaluate its performance.

## Materials and Methods

In this section, we outline the methodology employed to assess the performance of mathematical models against machine learning regression models. The mathematical model proposed by Obibuike and his team, focusing on a horizontal natural gas pipeline, serves as the cornerstone of our analysis.

The steps applied for the simulation of these models are as follows:

### Mathematical model (modified Weymouth equation) [2, 16]

Considering a natural horizontal gas pipeline, Obibuike developed a mathematical equation with several assumptions as stated in his research. The equations developed has been applied in the prediction of the pressure at which leak can occur and the leak location [15, 16]. The Leak pressure is given by the eqn. 1 while the location of the leak is defined by eqn. 2 respectively.

Where;
$P_{leak}$ = Leak Pressure
$X_{leak}$ = Leak Location
$P_{in}$ = Gas pipeline inlet pressure
$P_{out}$ = Gas pipeline outlet pressure
$q_{in}$ = Gas inlet flowrate
$q_{out}$ = Gas outlet flowrate
$X_{leak}$ = Leak location
L = length of pipeline

These parameters of the gas pipeline were used to develop the leak detection mathematical model modified from the Weymouth equation. In comparison to the machine learning models, the simulation of this model was done with python programming language for accurate evaluation and validation of the model's performance with the steps shown in Figure 1 below.

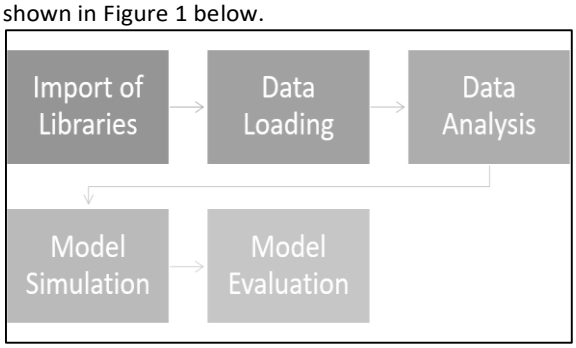



**Figure 1** Steps to simulation of the mathematical model on python program

Python programming was applied for the validation of the model with codes executed on the jupyter notebook.

1. **Import of libraries:** The python libraries used for the execution of the simulation process of the mathematical model are NumPy – for Numerical computation, Pandas – for data manipulation and matplotlib for visualization of the trends which exist in the dataset. These libraries were adopted for better experience and for good evaluation of the performance of the model.
2. **Data loading:** A natural gas dataset which comprises of the parameters which are dominant in the model was loaded into the notebook which was in form of a comma separated value (.csv), the loading of this dataset was made possible through the use of the library called pandas which allows creation of tables with rows and columns in python.
3. **Data analysis**: Data analysis was carried out on the dataset to ensure the dataset was void of errors and nulls values. During this several statistical assumptions were made and established to ensure the dataset is in good shape and is responsive to the codes and the model.
4. **Model simulation:** In Model simulation, the leak pressure model (equation) and the leak location model (equation) were coded into the python program language and given variable names respectively for which were made ready for evaluation.
5. **Model evaluation:** In model evaluation, the model coded was subjected to test on the dataset and predictions allowed to be made on the datasets for which it was fed and evaluated with metrics on

from the Sklearn library. The metrics used are; statistical metrics such as Mean Absolute Error, Mean Square Error, Root Mean Square Error, etc. With this, the performance of the model is displayed below.

### Machine learning model development

In machine learning, the same dataset was used to train three different machine learning models for which the performances were measured and evaluated to compare with that from the mathematical models. The procedure to training the machine learning models are itemized below which comprises of five major steps shown in Figure 2.

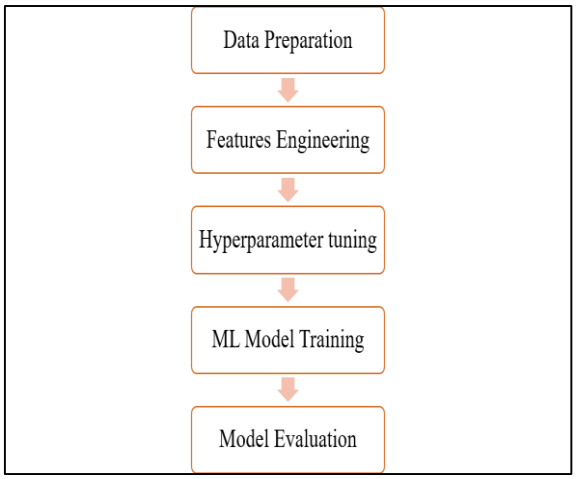



**Figure 2** Steps for training a machine learning model

1. **Data preparation:** This involves the manipulation and wrangling of the dataset to suit the purpose for which is to be used for with respect to the nature of the dataset. It includes; checking for missing values, outlier's treatment, exploring for other errors and cleaning the dataset to be void of these unconformities.
2. **Feature engineering:** Since machine learning models are intelligent models and performs better when the structure of a dataset is coherent and is in line which the expectation of the models, thus there is need for the engineering of the dataset which cut across the extraction of features which are relevant to the modelling process and the scaling of the dataset to ensure they have the same standard deviation for improved performance during evaluation.
3. **Hyperparameter tuning:** This process involves the selection of the hyperparameters for which will result in excellent performance. The GridSearchCv algorithm from Sklearn was applied in the selection process of these parameters.
4. **ML model training:** This is the most important step where the dataset being divided into the train and the test data of 80 to 20 percentage ratio for which the train and the test data were obtained respectively. The train dataset was used for the training of the ML model for the ML model was expected to study the patterns inherent the dataset and predictions made on the test dataset. In this study, three models were trained which include two ensemble models (random forest regressor and the voting regressor) and a booster model (XtraGradient Boost regressor).
5. **Model evaluation:** In model evaluation, the statistical error metrics were applied to measure the performance of the model. The same metrics used in the mathematical models were applied as well to ensure there is reduction in bias during comparison.

## Results and Discussion

The result is done for the two different aspects, the result from the mathematical modelling and the result from the machine learning modelling.

### Mathematical model results

After the simulation of the mathematical models and predictions made for the leak pressure and the leak location of a gas pipeline, the result of the statistical metrics is presented below

**Table 1** Mathematical model performance

| Mean Absolute Error | Mean Square Error | Root Mean Square Error | R-Square Score |
|---|---|---|---|
| 40.1475 | 2419.3790 | 49.1872 | 0.8877 |

Figure 3 Statistical Error plot for Mathematical model. This implies that the performance of the mathematical model was observed to be 88.87% with the errors scores as presented in Table 1 and Figure 3.


**Figure 3** Statistical error plot for mathematical model.

### Machine learning model training results

After the testing and the evaluation of the machine learning models trained, the performance of the models is presented below with visualizations. From Table 2, it is experienced that the performance of the machine learning models trained recorded 90+% accuracy both on the train and test dataset which implies better performance compared to the mathematical models. Trend analysis was performed to observe the trend and the path for which the mathematical and the machine learning models predicted the leaks experience in the gas pipelines. The results are presented in Figures 4, 5, 6, and 7 for mathematical model, random forest regressor, XGBoost regressor, and voting regressor model respectively.

**Table 2** Machine learning performance result.

| ML Models | Mean Absolute Error | Mean Square Error | Root Mean Square | Accuracy on train data | Accuracy on test data | R-Square Score |
|---|---|---|---|---|---|---|
| **Random Forest Regressor** | 0.4379 | 1.1460 | 1.0705 | 0.9962 | 0.9094 | 0.9094 |
| **XGBoost Regressor** | 0.5966 | 1.1306 | 1.0633 | 0.9998 | 0.9107 | 0.9107 |
| **Voting Regressor** | 0.4948 | 1.0430 | 1.0213 | 0.9990 | 0.9176 | 0.9176 |

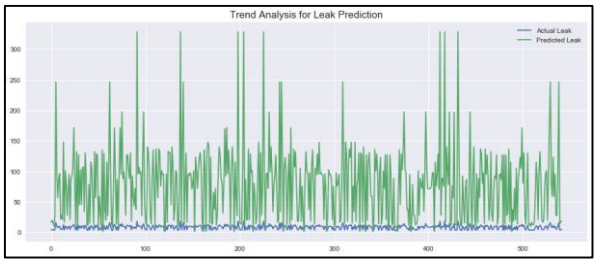



**Figure 4** Trend analysis for mathematical model prediction

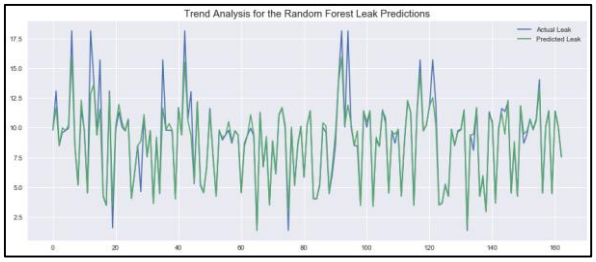



**Figure 5** Trend analysis for random forest regressor model

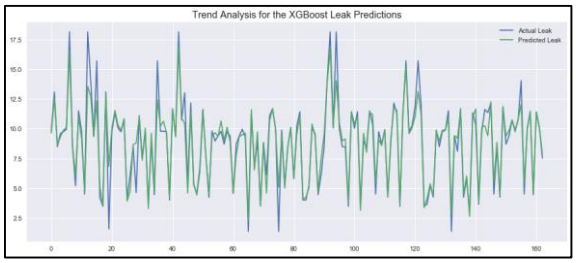



**Figure 6** Trend analysis for XGBoost regressor model

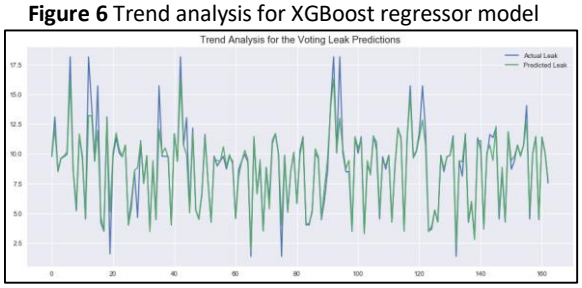



**Figure 7** Trend analysis for voting regressor model

From the trend analysis results shown in Figures above, the trend between the actual leak location and the prediction leak location are depicted by the blue and the green colour lines respectively. The analysis shows that the machine learning models were able to predict accurately the leak location with above 90% accuracy more than the mathematical models.

### Limitations of mathematical models

Mathematical models have several limitations which cut across the flowing:

1. Poor performance when evaluated on large dataset
2. Predictions from mathematical models are not trusted due to its poor performance score.
3. Prone to errors during simulation or when coding as it requires the coding of a long equation into the Integrated Development Environment (IDE).

These limitations were experienced during the course of this study for which machine learning models accounted for these limitations by providing an excellent performance, with speed in handling large datasets and reduced error of less than 3%.

## Conclusions

In conclusion, our comparative analysis between mathematical models and machine learning models unequivocally demonstrates the superior performance of the latter in detecting leaks in gas pipelines. The machine learning regression models, including the random forest regressor, XGBoost regressor, and voting regressor, consistently outperformed the mathematical model, boasting accuracies exceeding 90% with minimal errors. Conversely, the mathematical model exhibited lower, hovering around 80%, coupled with higher error rates.

These findings underscore the efficacy of machine learning models, particularly in scenarios involving large datasets, where quick response and accurate detection are paramount. As such, we advocate for the integration of machine learning regression models alongside existing sensor devices for enhanced leak detection capabilities in gas pipelines. By leveraging advanced technologies like machine learning, we can not only mitigate the environmental and economic repercussions of pipeline leaks but also pave the way for more efficient and reliable infrastructure management practices.

## Funding Sources

This research received no external funding

## Conflicts of Interest

There are no conflicts to declare.

## References

[1] Chis, A. P. T. (2007). Pipeline Leak Detection Techniques. Annals Computer Science Series, V(1).

[2] Obibuike, U. J., Kerunwa, A., Udechukwu, M., Eluagu, R. C., Igbojionu, A. C., & Ekwueme, S. T. (2020). Mathematical approach to determination of the pressure at the point of leak in natural gas pipeline. International Journal of Oil, Gas and Coal Engineering, 8(1), 22–27.

[3] Olawuyi, D. S., & Tubodenyefa, Z. (2018). Review of the environmental guidelines and standards for the

petroleum industry in Nigeria (EGASPIN). OGEES Institute.

[4] Nicola, M., Nicola, C., Vintilă, A., Hurezeanu, I., & Du, M. (2018). Pipeline leakage detection by means of acoustic emission technique using cross-correlation function. Journal of Mechanical Engineering and Automation, 8(2), 59–67.

[5] Badiola, X., & Vicente, J. (2021). Prediction using acoustic emission signals. Sensors, 1–16.

[6] Ossai, C. I. (2019). A data-driven machine learning approach for corrosion risk assessment — A comparative study. Big Data and Cognitive Computing.

[7] Fedushko, S. (2020). Predicting pupil's successfulness factors using machine learning algorithms and mathematical modelling methods. Springer International Publishing. https://doi.org/10.1007/978-3-030-16621-2

[8] Fokas, A. S., Dikaios, N., & Kastis, G. A. (2020). Mathematical models and deep learning for predicting the number of individuals reported to be infected with SARS-CoV-2. Royal Society Publishing.

[9] Saleem, F., Al-ghamdi, A. S. A., Alassafi, M. O., & Alghamdi, S. A. (2022). Machine learning, deep learning, and mathematical models to analyze forecasting and epidemiology of COVID-19: A systematic literature review. International Journal of Environmental Research and Public Health.

[10] Pinheiro, R., Vasconcelos, F., Maria, A., & Fileti, F. (2020). Machine learning and acoustic method applied to leak detection and location in low-pressure gas pipelines. Clean Technologies and Environmental Policy. https://doi.org/10.1007/s10098-019-01805-x

[11] Robinson, S. (2018). K-nearest neighbors algorithm in Python and Scikit-Learn. Stack Abuse. Stack Abuse. https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn/

[12] Boso, D. P., Di, D., Santagiuliana, R., Decuzzi, P., & Schrefler, B. A. (2020). Drug delivery: Experiments, mathematical modelling, and machine learning. Computers in Biology and Medicine, 123, 103820. https://doi.org/10.1016/j.compbiomed.2020.103820

[13] Dutta, N., Subramaniam, U., & Padmanaban, S. (2019). Mathematical models of classification algorithm of machine learning. International Meeting on Advanced Technologies in Energy and Electrical Engineering, 2018–2019.

[14] Kim, J., Chae, M., Han, J., Park, S., & Lee, Y. (2021). The development of leak detection model in subsea gas pipeline using machine learning. Journal of Natural Gas Science and Engineering, 94, 104134. https://doi.org/10.1016/j.jngse.2021.104134

[15] Igbojionu, A. C., Obibuike, U. J., Udechukwu, M., Mbakaogu, C. D., & Ekwueme, S. T. (2020). Hydrocarbon spill management through leak localization in natural gas pipeline. International Journal of Oil, Gas and Coal Engineering, 8(6), 137–142. https://doi.org/10.11648/j.ogce.20200806.13

[16] Julian, O. U., Toochukwu, E. S., Princewill, O. N., Chinwuba, I. K., Michael, O. I., & Chemazu, I. A. (2019). Analytical model for the estimation of leak location in natural gas pipeline. International Journal of Oil, Gas and Coal Engineering, 7(4), 95–102. https://doi.org/10.11648/j.ogce.20190704.12

[17] Akinsete, O., & Oshingbesan, A. (2019). Leak detection in natural gas pipelines using intelligent models. In SPE-198738-MS, August, 5–7.

[18] Santos, R. B., Sousa, E. O. De, Silva, F. V., Cruz, S. L., & Fileti, A. M. F. (2014). Detection and on-line prediction of leak magnitude in a gas pipeline using an acoustic method and neural network data processing. Journal of Natural Gas Science and Engineering, 31(1), 145–153.

[19] Kampelopoulos, D. (2020). Machine learning model comparison for leak detection in noisy industrial pipelines. In 2020 9th International Conference on Modern Circuits and Systems Technologies (MOCAST).

[20] Chakraborty, D., & Elzarka, H. (2018). Advanced machine learning techniques for building performance simulation: A comparative analysis. Journal of Building Performance Simulation, 1493. https://doi.org/10.1080/19401493.2018.1498538