EasyChair Preprint

№ 5579

# Comparison of Machine Learning Classifiers for Predicting Water Main Failure

Mohammad Amini and Rebecca Dziedzic



May 20, 2021

CSCE 2021 Annual Conference
*Inspired by Nature – Inspiré par la Nature*

26-29 May 2021

# COMPARISON OF MACHINE LEARNING CLASSIFIERS FOR PREDICTING WATER MAIN FAILURE

Amini, M[1, 3], Dziedzic, R.[2]
[1] Master's Student, Concordia University, Canada
[2] Assistant Professor, Concordia University, Canada
[3] mohammad.amini@concordia.ca

**Abstract:** Many water utilities are currently struggling to manage their aging infrastructure. Water mains are a key component of water systems, as they convey drinking water to billions of end-users worldwide. However, as they are usually buried underground, their visual inspection and condition assessment can be cumbersome. Furthermore, water main failure may lead to significant challenges for utilities and end-users, such as service interruption, capacity reduction, as well as high replacement and rehabilitation costs. Accordingly, various researchers have sought to develop statistical methods to predict water main condition. Previous studies have developed models for single systems, applying a range of statistical and machine learning methods, from linear regression to artificial neural networks. The objective of the present study is to compare the applicability and accuracy of a few machine-learning algorithms, such as Random Forest, Logistic Regression and Decision Tree to predict whether a pipe is going to fail or not (Classification). Data from two Canadian municipalities has been collected (Saskatoon, Saskatchewan and Waterloo, Ontario). A number of features are taken into consideration, such as diameter, age, material, and the number of previous failures. The results show a moderate to high accuracy of classification models although in some cases the performance of models is relatively low. Thus, deeper data mining approaches with higher concentrations on the most influential attributes would increase the reliability of the models.

## 1 INTRODUCTION

Water networks are among the most essential infrastructure worldwide, as they convey potable water to billions of end-users. Water distribution networks are reported to make up approximately 80% of total expenditures associated with the water industry (Kleiner and Rajani, 2001). According to ASCE (2017), approximately 240,000 incidents occur annually in the US, leading to around 1$ trillion in rehabilitation backlog required to improve the condition of water network components. A recent 2018 study found that 16% of installed pipes were beyond their useful life whereas a similar study in 2012 found 8% were beyond their useful life, and many utilities are lacking adequate fund to replace all of them (Folkman, 2012, 2018). Moreover, in US and Canada, overall water main failure is reported to have surged between 2012 and 2018, from 11 to 14 Failures/year/100 mile, respectively (Folkman, S. 2018). It was also reported that the rate of failure for Cast Iron (CI) and Asbestos Cement (AC) pipes increased by 40% during the aforementioned 6-year period. It should be noted that these two types of pipes account for almost 41% of all installed pipe in US and Canada. Furthermore, in 2017, the ASCE report card was prepared and given grade D to drinking water infrastructure in the USA as opposed to D- in 2009. Canada Infrastructure Report Card (2016) also reported that 29% of potable water infrastructures in very poor, poor or fair condition with a cost of $60 billion to replace. This is comparable to the 25% found in the latest 2019 report ("Canada Infrastructure Report Card" 2019).

In Canada, 59% of pipes were reported to be less than 40 years old and only 9% above 80 ("Canada Infrastructure Report Card" 2019). However, if reinvestment is not increased in Canada, the condition of core infrastructure may worsen, increasing the cost and risk of service interruption

("Canada Infrastructure Report Card" 2016). Mirza (2007) found that Canada had an estimated $123 billion total infrastructure backlog, $31 billion of which was related to water and wastewater networks.. In 2016, the Canada Infrastructure Report Card reported that 24.2$ billion is needed to maintain the water network in Canada. Water main deterioration may lead to service interruption, decrease in hydraulic capacity in the network and declining the quality of water flowing within the network (Kleiner and Rajani, 2001). Confronting these consequences, water network agencies are striving to develop new strategies to tackle the challenges pertinent to this important infrastructure. This highlights the importance of predictive models to plan and enhance the more efficient rehabilitation/maintenance operations (Dawood et al. 2020).

In recent decades, many studies have been conducted in order to find an appropriate method to assess the condition of water distribution networks (Giraldo-González and Rodríguez, 2020). A variety of physical and statistical models, as well as data-driven and machine learning algorithms have been utilized to predict the deterioration process of water main. For instance, Artificial Neural Networks (Al-Barqawi and Zayed, 2008; Jafar et al., 2010), Gradient Boosting Algorithm (Snider and McBean, 2018) and Random Forest (Shirzad and Safari, 2019) have been used to evaluate the condition of water networks. Physical models, on the other hand, are more comprehensive, however, acquiring data required for these models may be costly, therefore these models are justifiable only for transmission networks, where cost of failure is significant (Giraldo-González and Rodríguez, 2020; Kleiner and Rajani, 2001). Statistical models, however, employ historical records to recognize an explicit failure patterns, and then utilize these patterns to predict the probability or rate of failure in the future (Kleiner and Rajani, 2001). These models can link the finding pattern to the pipe features such as age, diameter, material, etc. (Giraldo-González and Rodríguez, 2020). This study focuses on three types of classification models: Decision Tree, Random Forest, and Logistic Regression.

## 2 Literature Review

Since the deterioration of water mains is an intricate process, attempts to forecast the failure of water pipes focus primarily on statistical models (Lei, J. and Saegrov, S. 1998). Statistical models utilize historical failure data in order to define patterns that are assumed to continue in the future. Kleiner and Rajani (2001) categorized these models to deterministic and probabilistic models. Using different attributes associated with water pipes, these models may estimate the probability of failure, rate of failure, and age at first and subsequent failures (Kleiner and Rajani, 2001; Park et al., 2011). Thus, comprehensive data would, undoubtedly, increase the accuracy of the models (Kleiner and Rajani, 2001). Nevertheless, the development process of such models for evaluating the condition of water mains is quite complex since failures typically occur as a result of different independent variables (Dawood et al. 2019).

The present study focuses on statistical deterministic models, using machine learning classifiers. Deterministic models predict specific rate of failure, or pipe age at failure based on historical failure rates (Kleiner and Rajani, 2001). They even can predict whether a pipe failed or not by associating them to machine learning models. These models require pipes to be partitioned into homogeneous groups that have similar characteristics. Such features could be material, size, soil characteristics and pipe vintage. This partitioning, however, imposes a challenge on the analysis process. That is, creating homogeneous groups may lead to unduly small groups. Simultaneously, these groups should be large enough for the statistical analysis to be reliable (Kleiner and Rajani, 2001).

There are a variety of machine learning classifiers, among which Decision Tree, Random Forest and Logistic Regression are employed in this study. The application of these models in previous studies are briefly described below.

### 2.1 Decision Tree

Decision Tree (DT) can be used either for classification or regression problems. Harvey and McBean (2014) applied DT to the prediction of sewer pipe failure in Guelph, ON, and compared it to Support Vector Machines. In this case, DT showed a 77% higher accuracy. Syachrani et al. (2013) also compared this type of model to regression and neural networks. Again DT outperformed other methods in terms of accuracy. However, DT has not been applied frequently for water distribution failure prediction, being more common for sewer systems (Oliveira et al. 2007). A DT model forecasts target labels with application of some predictive rules that are shaped in a structure similar to a tree (Syachrani et al. 2013). Rule building initiates from the root of the

tree where the dataset is assigned. This process continues by roots split into branches, which are known as decision nodes. This splitting process will not typically stop until the detection of only one class in a node which is called leaf. Figure 1 depicts the concept of the DT algorithm (Syachrani et al. 2013).

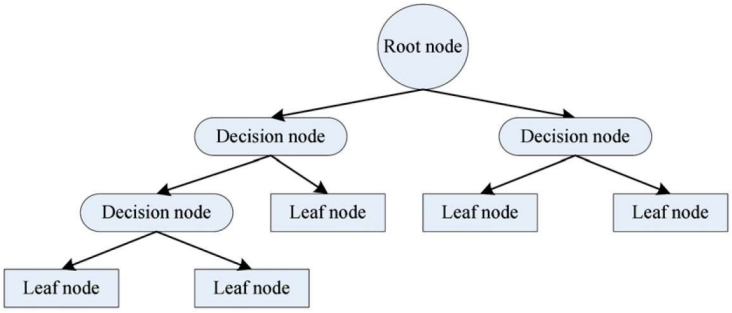


Figure 1 – Concept of decision tree algorithm (Syachrani et al. 2013)

### 2.2 Random Forests

Random Forest is a specific classification and regression algorithm based on the mixture of several decision trees (Breiman 2001). This algorithm has been used in several water-related studies (Chen et al., 2017; Shirzad and Safari, 2019; Zhu and Pierskalla, 2016). Sadler et al. (2018) reported that the traditional DT is prone to over fitting while applying it to the training data. The randomness of the Random Forest prevents over-fitting and leads to better model performance. In the conventional decision tree approach, the dataset is separated into smaller parts by using the best variable splitter, whereas Random Forest changes the splits while choosing random predictors. This makes Random Forest a more robust algorithm than DT. There is no assumption for Random Forest when splitting samples, and it clusters everything automatically (Vitorino et al. 2014).

### 2.3 Logistic Regression

Logistic Regression is a specific form of generalized linear model regression (Robles-Velasco et al. 2020; Vladeanu and Koo 2015). It can be used to calculate the probability of a sample pertaining to one class (Chang et al. 2019). Or in this case, can be used to classify binary values, such as broken and none broken pipes.

## 3 Methodology

### 3.1 Data cleaning

The first step in developing a statistical predictive model is to clean and prepare the data. Cleaning facilitates the subsequent modelling steps since many outliers and inconsistent entries are removed. Thus, the accuracy of the predictive models can be enhanced. The case study datasets include pipe diameter, pipe length, age at failure, month of failure, and material. In the present study, only pipes with a length of 200m or less were considered, as well as an age of 80 years or less, since less information is available for older breaks.

Data on water main and breaks is generally available in two separate datasets, an inventory and a main break register. These datasets were merged, allowing for the identification in the inventory of pipes that have failed or not.

In order to account for the impact of different monthly weather condition, the month at time of break was defined as a binary value (0 or 1) for each month. Furthermore, although there is a range of materials within the datasets, the most common materials were used in the analysis, Asbestos Cement (AC), Cast Iron (CI), Ductile Iron (DI), Polyvinyl Chloride (PVC) and Polyethylene (PE),

focusing on CI pipes. Each material was also treated as a binary value, and 0 or 1 was assigned to each material. Table 1 and

Table *2* which are provided in section 4 show the most important attributes that have been used in this analysis.

### 3.2 Classification of Broken and Non-Broken Pipes

Three different machine learning algorithms were applied to both datasets, Random Forest Classifier, Logistic Regression Classifier, and Decision Tree. Each model was applied to the Saskatoon and Waterloo data separately, which includes five types of materials; AC, CI, PVC, PE, and DI. The datasets were divided into training set (70%) and test set (30%). Accuracy was evaluated on the test set, as well with a 5-fold cross validation approach. Such an evaluation (Cross Validation) indicates accuracy among different portions of data. In addition to accuracy, precision, recall and F-1 score were also calculated. These classification metrics present the prediction power of the applied classification models.

Once the models were developed for each system, the Saskatoon model was tested on Waterloo data. This allows for the comparison of the models, and evaluation of their applicability under other conditions. There are many environmental factors that contribute to pipe failure, such as temperature, soil type, etc. which vary among different systems. Thus, Waterloo and Saskatoon have different characteristics within their water networks that are not currently included in the dataset.

Since homogeneity is reported to be significantly important while making a predictive model (Shamir and Howard 1979), Cast Iron (CI) pipes were also analyzed separately for Saskatoon and Waterloo.

## 4 Case Study Systems

The data for this analysis was collected from the cities of Saskatoon, Saskatchewan and Waterloo, Ontario. Water main inventories, as well as water main breaks were used to predict pipe failure. Main break data is available for 2000 to 2019. Saskatoon and Waterloo water networks consist of 1,188 KMs and 433 KMs of water mains, respectively. This includes a variety of materials such as Cast Iron (CI), Asbestos Cement (AC), Ductile Iron (DI), Polyvinyl Chloride (PVC), Polyethylene (PE), High-density polyethylene (HDPE) and Flexible Polyvinyl Chloride (FPVC). However, the final materials included in the analysis are Cast Iron (CI), Ductile Iron (DI), PVC, Polyethylene (PE) and Asbestos Cement. Table 3 shows the proportion of each material within both network after data cleaning. AC and CI pipes account for almost 48% of the Saskatoon network, 564 km. In Waterloo, CI pipe makes up around 30% of the network, 128 km. Thus, this report focuses primarily on Cast Iron pipes. The most important attributes employed in the analysis for both city of Saskatoon and Waterloo (Table 1 and Table 2).

There are also different pipes with different ages in the inventory. The distribution of age versus cumulative length of pipes for both networks is provided in the given figures (Figure a) and (Figure b). It should be noted that the age distribution has been narrowed down, in order to improve the accuracy of the model.

Table 1: Summary of input attributes for Waterloo

| Attribute | Count | Mean | STD | Min Value | Max Value |
|---|---|---|---|---|---|
| **Diameter (mm)** | 28869 | 200.84 | 64.31 | 40 | 600 |

| | | | | | |
|---|---|---|---|---|---|
| **Length (m)** | 28869 | 30.49 | 39.61 | 0.20 | 199.97 |
| **Age (years)** | 28869 | 27.37 | 19.61 | 0 | 80 |
| **Break Month: Jan – Dec (Binary Variable)** | 28869 | - | - | 0 | 1 |
| **PVC** | 28869 | 0.54 | 0.50 | 0 | 1 |
| **Asbestos Cement** | 28869 | 0.32 | 0.47 | 0 | 1 |
| **Cast Iron** | 28869 | 0.13 | 0.34 | 0 | 1 |
| **PE** | 28869 | 0.00 | 0.05 | 0 | 1 |
| **Ductile Iron** | 28869 | 0.00 | 0.05 | 0 | 1 |

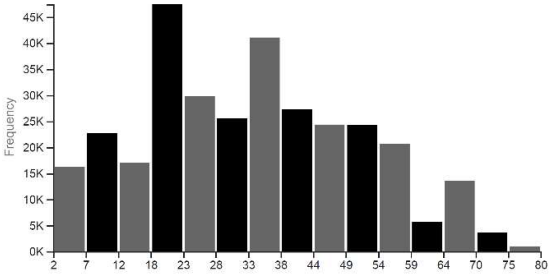


Figure 2a: Distribution of age versus cumulative length of pipe (Waterloo) - Age/Thousands Meters

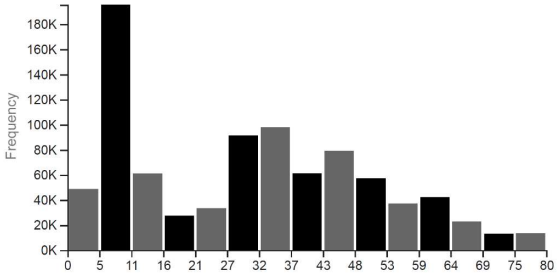


Figure 2b: Distribution of age versus cumulative length of pipe (Saskatoon) – Age/Thousands Meters

Table 2: Summary of input attributes for Waterloo

| Attribute | Count | Mean | STD | Min Value | Max Value |
|---|---|---|---|---|---|
| **Diameter (mm)** | 6884 | 202.06 | 70.15 | 25 | 150 |
| **Length (m)** | 6884 | 46.50 | 52.25 | 0.10 | 5.80 |
| **Age** | 6884 | 34.44 | 18.10 | 2 | 20 |
| **Break Month : Jan – Dec (Binary Variable)** | - | - | - | 0 | 1 |
| **PVC** | 6884 | 0.58 | 0.49 | 0 | 0 |
| **Asbestos Cement** | 6884 | 0.00 | 0.05 | 0 | 0 |
| **Cast Iron** | 6884 | 0.27 | 0.44 | 0 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| **PE** | 6884 | 0.00 | 0.02 | 0 | 0 |
| **Ductile Iron** | 6884 | 0.15 | 0.36 | 0 | 0 |

Table 3: Proportion of each type of material in both networks

| **Saskatoon** | | | **Waterloo** | | |
|---|---|---|---|---|---|
| **Material** | **Percentage** | **Length (km)** | **Material** | **Percentage** | **Length (km)** |
| Asbestos Cement (AC) | 38.12% | 335.5 | Asbestos Cement (AC) | 0.25% | 0.80 |
| Cast Iron (CI) | 15.02% | 132.2 | Cast Iron (CI) | 25.05% | 80.17 |
| Ductile Iron (DI) | 0.14% | 1.2 | Ductile Iron (DI) | 15.16% | 48.53 |
| Poly Ethylene (PE) | 0.13% | 1.1 | Poly Ethylene (PE) | 0.07% | 0.21 |
| PVC | 46.60% | 410.1 | PVC | 59.48% | 190.39 |
| Total | 100% | 880.07 | Total | 100% | 320.10 |

## 5 Results

### 5.1 Classification Metrics

Several primary classification metrics can be employed in order to analyze the reliability of classifiers. These include accuracy, precision, recall and F-1, and are calculated using a confusion matrix. The confusion matrix not only indicates the general accuracy of the model, but demonstrates the incorrect and correct prediction of a classifier. This matrix includes the number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) as defined in Table 4.

Table 4: Document Classification (Ikonomakis, et al. 2005)

| | |
|---|---|
| True Positive | True positive labels in the main dataset |
| True Negative | True negative labels in the main dataset |
| False Positive | False prediction of positive value |
| False Negative | False prediction of negative value |

Considering these terms, the following evaluation metrics can be calculated.

- ***Accuracy:*** This metric can be used to evaluate True results (negative and positive) compared to all results (TP, TN, FP and FN). It is straightforward to understand and typically used for evaluation of datasets including balanced classes (Negative or Positive value of target labels are relatively equal). When imbalanced classes are available in the dataset, accuracy alone could result in misinterpretation (Harvey and McBean 2014).

- ***Precision:*** This metric can be used to evaluate Positive results and the accuracy of positive prediction. When positive results are influential in the analysis, this metric would be helpful.

- ***Recall:*** This metric is employed for evaluation of actual positive classification. When the importance of positive target is important, recall can be used.

- ***F-1:*** In order to have better understanding of accuracy of the classifiers, precision and recall could be combined. This combination would result in F-1 which is a reliable indicator, showing the accuracy of classification models. This metric also can be used in a case that target label has imbalanced values. Table 5 shows the equations for each of these metrics.

Table 5: Evaluation metrics (Syachrani et al. 2013)

| | |
|---|---|
| **Accuracy** | (TP+TN) / (TP+FP+FN+TN) |
| **Precision (P)** | (TP)/(TP+FP) |
| **Recall (R)** | (TP)/(TP+FN) |
| **F-1** | 2 * ((P*R) / (P+R)) |

### 5.2 Classification Results

Evaluation metrics of the classification models applied to Saskatoon and Waterloo are provided in Table 6. The Random Forest classifier outperformed other models in both networks in the first step (dataset including different materials) with general accuracy of 77% and 97%, respectively. However, F-1 score for Waterloo and Saskatoon is higher for Logistic Regression model. Moreover, after applying cross validation to test the models, Logistic regression indicated better accuracy with 60% of accuracy and 83% recall. In waterloo, however, Logistic Regression did not show a significant improve after cross validation method. Random Forest found age and length to be the most important predictors with 43% and 41% contribution, respectively. However, Decision Tree considered age as the most important contributor with 75% weight. The significant difference between F-1 score in both case studies indicates the dependency of the predictive models on each specific site factors. It can be seen F-1 score shows higher accuracy for Saskatoon dataset. The analysis carried on with applying models to Cast Iron (CI) pipes in Saskatoon. Interestingly, with partitioning the dataset into specific material type, the accuracy of the models increased relatively. For instance, Random Forest Accuracy surged from 77% in the first step to 81%, and Decision Tree from 74% to 81%. In this step Random Forest and Decision Tree represented a better performance comparing to Logistic Regression. Recall and Precision are almost same for all three classifiers. In this case, F-1 score for Saskatoon is relatively higher, indicating the higher reliability of classifiers for Saskatoon network. In this preliminary study, age and length found to be the most important attributes affecting the prediction results. Overall, the results demonstrate somewhat reliability of machine learning classifiers for forecasting whether a pipe fails in the future. However, according to previous studies (Andreou S. A. 1986; Clark et al. 1982; Shamir and Howard 1979) many other factors may have significant impact on water main failure, which are not available in this study. These factors could be soil resistivity, temperature, soil type and other operational and environmental features. Additionally, the importance of partitioning data into homogenous groups can be noticed clearly from the result, as the accuracy of the models is clearly improved by grouping them by material. Decision Tree and Random Forest classifiers seem to be appropriate methods for evaluation and prediction.

Table 6: Comparison of classifiers – (Saskatoon - Waterloo)

| **Classification Models Comparison - (Saskatoon - All types of materials)** | | | | | |
|---|---|---|---|---|---|
| **Model** | **Accuracy** | **CV Accuracy** | **CV Precision** | **CV Recall** | **CV F1** |
| Random Forest | 77% | 56% | 65% | 77% | 71% |
| Logistic Regression | 66% | 60% | 67% | 83% | 74% |
| Decision Tree | 74% | 57% | 65% | 79% | 71% |
| **Classification Models Comparison - (Saskatoon - Cast Iron Pipes)** | | | | | |
| **Model** | **Accuracy** | **CV Accuracy** | **CV Precision** | **CV Recall** | **CV F1** |
| Random Forest | 81% | 69% | 75% | 85% | 80% |

| | | | | | |
|---|---|---|---|---|---|
| Logistic Regression | 69% | 68% | 74% | 85% | 79% |
| Decision Tree | 81% | 70% | 76% | 85% | 80% |

**Classification Models Comparison - (Waterloo - All types of materials)**

| Model | Accuracy | CV Accuracy | CV Precision | CV Recall | CV F1 |
|---|---|---|---|---|---|
| Random Forest | 97% | 94% | 73% | 40% | 38% |
| Logistic Regression | 96% | 95% | 78% | 34% | 41% |
| Decision Tree | 96% | 94% | 62% | 35% | 39% |

**Classification Models Comparison - (Waterloo - Cast Iron Pipes)**

| Model | Accuracy | CV Accuracy | CV Precision | CV Recall | CV F1 |
|---|---|---|---|---|---|
| Random Forest | 92% | 92% | 77% | 59% | 65% |
| Logistic Regression | 92% | 90% | 73% | 49% | 58% |
| Decision Tree | 93% | 90% | 73% | 53% | 60% |

In order to evaluate the performance of classifiers, same process has been done for Waterloo, and the results are provided in Table 6. Although the general accuracy of the models is relatively high in the first step (dataset including different materials), after cross validation, the prediction power of models decreased significantly. As it can be seen, recall and F-1 scores for all types of material are not satisfactory. However, after partitioning data and applying the same models to only CI pipes, the accuracy of models in prediction somewhat improved. This again emphasizes the importance of partitioning pipes in homogeneous classes. In this section of analysis length and age with 30% and 36% contribution – according to feature importance analysis in open source tools - seemed to be the most important factors that may affect the prediction results. However, this results are not finalized and they are site-specific. Hence, more analysis is required to prove the reliability of the models. As it can be seen F-1 score for Waterloo after cross-validation method increased, which emphasize the importance of partitioning data into homogeneous groups.

## 6 Summary and Conclusions

The present study focused on failure prediction of water pipes considering pipe age, length, material, and month of failure. Results show that classifiers can provide useful and moderately accurate predictions of pipe failure. However, there are many other attributes that may contribute to failure and could be taken into consideration to increase the reliability of the model. Other significant factors could include soil type, previous rate of failure, temperature, and soil resistivity. Next steps of this project involve including other variables as well as other Canadian cities to the analysis.

## 7 References


Al-Barqawi, Hassan, and Tarek Zayed. 2008. “Infrastructure Management: Integrated AHP/ANN Model to Evaluate Municipal Water Mains’ Performance.” *Journal of Infrastructure Systems* 14 (4): 305–18. https://doi.org/10.1061/(ASCE)1076-0342(2008)14:4(305).

Andreou S. A. 1986. “Predictive Models for Pipe Break Failures and Their Implications on Maintenance Planning Strategies for Deteriorating Water Distribution Systems.”

ASCE. 2017. “ASCE Drinking Water Report Card.”

Chang, Minwoo, Marc Maguire, and Yan Sun. 2019. “Stochastic Modeling of Bridge Deterioration Using Classification Tree and Logistic Regression.” *Journal of Infrastructure Systems* 25 (1): 04018041. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000466.

Chen, Guoqiang, Tianyu Long, Jiangong Xiong, and Yun Bai. 2017. “Multiple Random Forests Modelling for Urban Water Consumption Forecasting.” *Water Resources Management* 31 (15): 4715–29. https://doi.org/10.1007/s11269-017-1774-7.

Clark et al. 1982. “Water Distribution Systems: A Spatial and Cost Evaluation.”

Dawood, Thikra, Emad Elwakil, Hector Mayol Novoa, and José Fernando Gárate Delgado. 2019. "Pipe Failure Prediction and Risk Modeling in Water Distribution Networks: A Critical Review." *Science and Technology*, 7.

Dawood, Thikra, Emad Elwakil, Hector Mayol Novoa, and José Fernando Gárate Delgado. 2020. "Water Pipe Failure Prediction and Risk Models: State-of-the-Art Review." *Canadian Journal of Civil Engineering* 47 (10): 1117–27. https://doi.org/10.1139/cjce-2019-0481.

Giraldo-González, Mónica Marcela, and Juan Pablo Rodríguez. 2020. "Comparison of Statistical and Machine Learning Models for Pipe Failure Modeling in Water Distribution Networks." *Water* 12 (4): 1153. https://doi.org/10.3390/w12041153.

Harvey, Robert Richard, and Edward Arthur McBean. 2014. "Comparing the Utility of Decision Trees and Support Vector Machines When Planning Inspections of Linear Sewer Infrastructure." *Journal of Hydroinformatics* 16 (6): 1265–79. https://doi.org/10.2166/hydro.2014.007.

Ikonomakis, M, S Kotsiantis, and V Tampakas. 2005. "Text Classification Using Machine Learning Techniques," 10.

Jafar, Raed, Isam Shahrour, and Ilan Juran. 2010. "Application of Artificial Neural Networks (ANN) to Model the Failure of Urban Water Mains." *Mathematical and Computer Modelling* 51 (9–10): 1170–80. https://doi.org/10.1016/j.mcm.2009.12.033.

Kleiner, and Rajani. 2001. "Comprehensive Review of Structural Deterioration of Water Mains: Statistical Models." *Urban Water* 3 (3): 131–50. https://doi.org/10.1016/S1462-0758(01)00033-4.

Lei, J. and Saegrov, S. 1998. "Statistical Approach for Describing Failures and Lifetimes of Water Mains," 9.

Oliveira, D., W. Guo, L. Soibelman, and James H. Garrett, Jr. 2007. "Spatial Data Management and Analysis in Sewer Systems' Condition Assessment: An Overview." In *Computing in Civil Engineering (2007)*, 391–98. Pittsburgh, Pennsylvania, United States: American Society of Civil Engineers. https://doi.org/10.1061/40937(261)48.

Park, Hwandon Jun, Newland Agbenowosi, Bong Jae Kim, and Kiyoung Lim. 2011. "The Proportional Hazards Modeling of Water Main Failure Data Incorporating the Time-Dependent Effects of Covariates." *Water Resources Management* 25 (1): 1–19. https://doi.org/10.1007/s11269-010-9684-y.

Robles-Velasco, Alicia, Pablo Cortés, Jesús Muñuzuri, and Luis Onieva. 2020. "Prediction of Pipe Failures in Water Supply Networks Using Logistic Regression and Support Vector Classification." *Reliability Engineering & System Safety* 196 (April): 106754. https://doi.org/10.1016/j.ress.2019.106754.

Roiger, Richard J. 2017. "Basic Data Mining Techniques." In *Data Mining*, by Richard J. Roiger, 2nd ed., 63–102. Chapman and Hall/CRC. https://doi.org/10.1201/9781315382586-3.

Sadler, J.M., J.L. Goodall, M.M. Morsy, and K. Spencer. 2018. "Modeling Urban Coastal Flood Severity from Crowd-Sourced Flood Reports Using Poisson Regression and Random Forest." *Journal of Hydrology* 559 (April): 43–55. https://doi.org/10.1016/j.jhydrol.2018.01.044.

Saeed Mirza. 2007. *Danger Ahead: The Coming Collapse of Canada's Municipal Infrastructure*. Ottawa, Ont.: Federation of Canadian Municipalities. https://www.deslibris.ca/ID/250220.

Shamir and Howard. 1979. "An Analytic Approach to Scheduling Pipe Replacement." *Journal - American Water Works Association* 71 (5): 248–58. https://doi.org/10.1002/j.1551-8833.1979.tb04345.x.

Shirzad, Akbar, and Mir Jafar Sadegh Safari. 2019. "Pipe Failure Rate Prediction in Water Distribution Networks Using Multivariate Adaptive Regression Splines and Random Forest Techniques." *Urban Water Journal* 16 (9): 653–61. https://doi.org/10.1080/1573062X.2020.1713384.

Snider, Brett, and Edward A McBean. 2018. "IMPROVING TIME-TO-FAILURE PREDICTIONS FOR WATER DISTRIBUTION SYSTEMS USING GRADIENT BOOSTING ALGORITHM," 8.

Steven Folkman. 2012. "Water Main Break Rates In the USA and Canada A Comprehensive Study," 28.

Steven Folkman. 2018. "Water Main Break Rates In the USA and Canada: A Comprehensive Study," 48.

Syachrani, Syadaruddin, Hyung Seok "David" Jeong, and Colin S. Chung. 2013. "Decision Tree–Based Deterioration Model for Buried Wastewater Pipelines." *Journal of Performance of Constructed Facilities* 27 (5): 633–45. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000349.

Vitorino, D., S.T. Coelho, P. Santos, S. Sheets, B. Jurkovac, and C. Amado. 2014. "A Random Forest Algorithm Applied to Condition-Based Wastewater Deterioration Modeling and Forecasting." *Procedia Engineering* 89: 401–10. https://doi.org/10.1016/j.proeng.2014.11.205.

Vladeanu, Greta Julia, and Dan D. Koo. 2015. “A Comparison Study of Water Pipe Failure Prediction Models Using Weibull Distribution and Binary Logistic Regression.” In *Pipelines 2015*, 1590–1601. Baltimore, Maryland: American Society of Civil Engineers. https://doi.org/10.1061/9780784479360.146.

Zhu, Junfeng, and William P. Pierskalla. 2016. “Applying a Weighted Random Forests Method to Extract Karst Sinkholes from LiDAR Data.” *Journal of Hydrology* 533 (February): 343–52. https://doi.org/10.1016/j.jhydrol.2015.12.012.