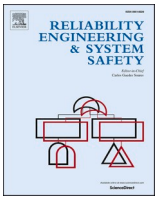
Reliability Engineering and System Safety 219 (2022) 108185

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress



Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors

Xudong Fan [a], Xiaowei Wang [b], Xijin Zhang [c], P.E.F. ASCE Xiong (Bill) Yu [d,*]

[a] Graduate Research Assistant, Department of Civil and Environmental Engineering, Case Western Reserve University, 2104 Adelbert Road, Bingham 248, Cleveland, OH 44106-7201, US
[b] Postdoctoral Scholar. Department of Civil and Environmental Engineering, Case Western Reserve University, 2104 Adelbert Road, Bingham 248, Cleveland, OH 44106-7201, US
[c] Graduate Research Assistant, Department of Civil and Environmental Engineering, Case Western Reserve University, 2104 Adelbert Road, Bingham 249, Cleveland, OH 44106-7201, US
[d] Opal J. and Richard A Vanderhoof Professor and Chair, Department of Civil and Environmental Engineering, Case Western Reserve University, 2104 Adelbert Road, Bingham 237, Cleveland, OH 44106-7201, US

ARTICLE INFO

Keywords:
Multi-source data aggregation
Machine learning
Water supply network
Pipe failure prediction

ABSTRACT

Underground water pipes deteriorate under the influence of various physical, mechanical, environmental, and social factors. Reliable pipe failure prediction is essential for a proactive management strategy of the water supply network (WSN), which is challenging for the conventional physics-based model. This study applied data-driven machine learning (ML) models to predict water pipe failures by leveraging the historical maintenance data heritage of a large water supply network. A multi-source data-aggregation framework was firstly established to integrate various contributing factors to underground pipe deterioration. The framework defined criteria for the integration of various data sources including the historical pipe break dataset, soil type dataset, topographical dataset, census dataset, and climate dataset. Based on the data, five ML algorithms, including LightGBM, Artificial Neural Network, Logistic Regression, K-Nearest Neighbors, and Support Vector Classification are developed for pipe failure prediction. LightGBM was found to achieve the best performance. The relative importance of major contributing factors on the water pipe failures was analyzed. Interestingly, the socioeconomic factors of a community are found to affect the probability of pipe failures. This study indicates that data-driven analysis that integrates the Machine Learning (ML) techniques and the proposed data integration framework has the potential to support reliable decision-making in WSN management.

1. Introduction

Providing a reliable and safe water supply is crucial for water supply network (WSN) management. Water distribution pipes are the key components of a WSN for delivering water from water treatment plants to customers. The buried water pipes are subjected to deterioration, this is especially severe for US urban municipalities where some of the pipes were laid down in the 19th century [1]. More than 700 water mains break every day in Canada and the USA, wasting approximately 2 trillion gallons of drinking water annually [2]. The failure of water pipes can cause significant financial losses with associated societal or environmental impacts. According to the American Water Work Association, the replacement costs of the existing WSNs in the US, combined with their projected expansions, would cost more than $1 trillion over the next couple of decades [3]. These devastating issues press for reliable pipe failure prediction models to institute preventative maintenance for loss reduction as well as a management practice for sustainable improvement.

Identifying the key factors (i.e., input variables) that influence the pipe failure is a key task in developing reliable prediction models. Over the past decades, various factors that could lead to pipe breaks have been assessed by experimental tests, finite element simulations, and historical data analysis [4,5]. According to a recent review, these factors can be generally grouped into three types according to their attributes in WSN, i.e. *physical, operational*, and *environmental* [6]. The most widely considered *physical* factors include the pipe age, length, material, and diameter. For example, Kettle and Goulter identified the relationship between the pipe diameter and the break probability using statistical

* Corresponding author.
E-mail addresses: xxf121@case.edu (X. Fan), xwang@case.edu (X. Wang), xxz677@case.edu (X. Zhang), xxy21@case.edu (P.E.F. ASCE Xiong (Bill) Yu).

https://doi.org/10.1016/j.ress.2021.108185
Received 21 July 2021; Received in revised form 24 October 2021; Accepted 26 October 2021
Available online 8 November 2021
0951-8320/

analysis [7]. Yamijala et al. demonstrated that pipes with longer lengths will suffer more failures [8]. The most widely examined *operational* factor is the number of times that previous failures occurred [9–11]. These studies indicate that a larger number of previous failures along a pipe is often associated with a higher probability of failure. In addition, water pressure is another common operational factor for pipes in the WSNs [12,13]. A positive correlation between internal water pressure and the probability of breaks in concrete and metallic pipes has been observed [13]. *Environmental* factors could also contribute to pipe failures. These factors include traffic loads, soil types, and climate [6]. Also, most of these factors often have high levels of uncertainty [14,15]. Previous studies identified the influence of different climate factors such as temperature and precipitation on pipe failures [9,16,17]. The results indicated that larger temperature fluctuations could increase pipe failure probability. These three types of factors interplay and there is a need to better understand the interplay between these contributing factors and pipe failure probabilities. Besides the above-mentioned factors, it has been increasingly realized that the interactions with other types of factors such as socioeconomic factors should be also considered in WSNs' performance prediction. For example, recent studies on community reliability and resilience considered the influence of infrastructure failure and census data [18]. However, such factors have rarely been considered in the existing pipe failure prediction model. Meanwhile, besides developing reliable and efficient tools for pipe failure evaluation, stakeholders are highly interested in understanding the mechanisms of the most important factors in pipe failures, which allows instituting informed decisions for resource allocation. Although previous studies have examined the impacts of different factors on pipe failure probabilities [19], the relative importance (i.e., degree of the impacts) is yet to be well understood. Therefore, the interpretability consideration is also very important in developing a pipe failure prediction model.

The existing methods for pipe failure prediction generally fall into three categories, i.e., *physics-based models, statistical models*, and *machine learning models* [20]. The advantages and limitations of each method are briefly reviewed here. *Physical-based models* consider physical factors with empirical or semi-empirical feature equations that compare the allowable strength of a pipe to the real-world loading [21–23]. Then, the failure probability of a pipe can be determined by comparing the pipe's remaining strength and its loading via a sampling method, such as the Monte Carlo simulation [24,25]. Although physics-based models could intuitively indicate the different contributions of the factors considered [26], this method is usually computationally intensive for the entire WSN system. It is because hundreds or even thousands of pipes are contained in a WSN and each requires an enormous number of samplings for the failure probability estimation [27]. Moreover, some of the popular physics-based models are too conservative, such as the B31G model [26]. By contrast, *Statistical models* are cost-efficient particularly for specific areas with sufficient historical recording data. Previous studies have reported the applicability of various statistical models for pipe failure prediction [28–31]. The statistical models typically used statistical equations (such as the time-exponential model, linear regression model, and the Poisson-process model) to model time-dependent breakage prediction models [32,33]. Recent studies also used the Bayesian networks for the pipe failure inference [34,35]. Under the assumption that the failure pattern remains consistent in the future, these models can be further used for failure prediction. However, the statistical models can only consider limited physical factors without revealing their physical relationships to pipe failures. Recently, data-driven *machine learning (ML) models* have become an emerging method for the prediction of pipe failures, leveraging the increasing amount of data available. Artificial neural networks (ANN), neuro-fuzzy systems, Logistic regression, and genetic algorithms are some of the most popular approaches to find the complex relationship between pipe failure and different variables [36–41]. Although these ML-based approaches can often achieve promising computation efficiency and accuracy, they are often criticized as 'black boxes' that lack or have limited interpretability. To overcome this limitation, some studies have used more explainable ML algorithms such as tree-based algorithms and Logistic regression algorithm. However, these algorithms may not achieve satisfactory accuracy for water pipe break prediction, based on the knowledge and experience of the authors. Moreover, although some of the ML algorithms have been used for pipe failure prediction, there is still a lack of systematic comparison among different types of ML algorithms in this application domain. In summary, even with valuable datasets maintained by water agencies alongside the WSN management, accurate prediction and reasonable interpretation of pipe failure probability is still challenging and thereby needs more work.

In light of the abovementioned research gaps, this study aims to propose a multi-source data-aggregation framework for an interpretable ML model that facilitates high-fidelity and efficient prediction of pipe failures in water supply networks. The proposed framework was applied on a large WSN dataset that contains more than 5300 miles (8529 km) of water pipe, which represents the largest analyzed dataset so far based on the authors' knowledge. Multiple sources including physical, operational, environmental, topographical, and census data are considered in the data preparation stage. The interpretation results not only fit previous studies but also indicate the important impact of socioeconomic factors. Although the factors' impact may vary in different WSNs, the proposed analysis framework can be easily adopted by different WSN management agencies.

2. Background of machine learning models for pipe failure prediction

A common approach for pipe failure prediction is treating this problem as a classification problem, i.e., classify a pipe as either broken or intact by the given variables. The current supervised ML algorithms for classification tasks can be generally divided into five main categories [42], i.e., logic-based algorithms, perceptron-based techniques, statistical learning algorithms, instance-based learning algorithms, and support vector machines. Although previous studies have used different ML algorithms to classify if a pipe will break or not based on the array of input conditions for the pipe, there is a lack of a comprehensive comparison between different types of ML models. To efficiently compare the performance of these ML categories for pipe break predictions, five popular ML algorithms are selected as the representative of each category, i.e., the LightGBM algorithm, the Artificial Neural Network (ANN), logistic regression, k-nearest neighbors (kNN), and support vector classification (SVC). The objective of the ML model is to classify each pipe into either intact or broken categories based on the observed input factors. A brief description of these ML algorithms is provided below.

2.1. LightGBM

LightGBM is a gradient boosting framework that belongs to logic-based classification algorithms. The LightGBM involves three features: Gradient-based One-Side Sampling (GOSS), Exclusive Feature Bundling (EFB), and histogram and leaf-wise tree growth strategy. More specifically, the GOSS is a sampling method by keeping all instances with large gradients and perform random sampling on the instances with small gradient. A constant multiplier for the data instances with small gradients is used to compensate the data distribution changing during the sampling process. Hence the final performance is improved by using these two strategies. The EFB method aims to increase the computational efficiency by dividing the features into a smaller number of bundles. LightGBM uses EFB together with histogram algorithms to efficiently deal with categorical features. Therefore, the categorical features do not need to be represented with the traditional one-hot encoding, which is a significant benefit especially when the categorical feature contains lots of unique values. Detailed theories and information can be found in [43].

2.2. Artificial neural network (ANN)

ANN is a widely used perceptron that mimics the working of human brain in processing a combination of stimuli into an output [44]. The key components of an ANN architecture are the input layer, hidden layers and output layer . The weight coefficients for neurons of each layer are iteratively tuned to mitigate the error between the output and target values during the training process. For each neuron, it computes a weighted sum of its inputs and generates an output with an activation function. Mathematically, for each neuron, its relationship between the inputs and output can be represented by Eq. (1).

$$\theta = f\left(\sum_{r=1}^{k} w_r x_r + b\right) \tag{1}$$

where θ is the output of each neuron, $f(\cdot)$ is the activation function, w_r is the weight of x_r and b is the bias.

2.3. Logistic regression (LR)

LR is one of the statistical learning algorithms by fitting samples into a logistic function. LR has been widely used in engineering areas because (1) it is an explainable ML algorithm since the weight for each factor is available after training, and (2) it assigned a value between 0 and 1 to each sample for a classification problem, which can be interpreted as the classification probability. Mathematically, LR is formulated as Eq. (2) [45].

$$p = \frac{1}{1 + e^{-\left(w_0 + \sum_{i=1}^{m} w_i x_i\right)}} \tag{2}$$

where p is the output of each sample; x_i is the vector sample with the i^{th} feature; w_i is the weight of the i^{th} feature that will be tuned during the training process; and w_0 is the constant bias. As can be observed from the equation LR cannot handle the categorical variables directly. Therefore, converting methods such as one-hot-encoder is required. Once the weights are determined, the classification result, y, of each sample can be achieved by Eq. (3), in which the threshold is usually set as 0.5 for a binary classification problem.

$$= \begin{cases} -1 \ if \ p \ \leq threshold \\ 1 \ if \ p > threshold \end{cases} \tag{3}$$

2.4. k-nearest neighbors (kNN)

kNN is one of the most classical and simplest methods for pattern classification [46]. It assumes the instances belongs to the same class exist in close proximity. The performance of kNN heavily depends on the way that the distances are computed between the instances. Euclidean distance is the most commonly used distance metric. The Euclidean distance between samples x_i and x_j is defined as Eq. (4).

$$d(x_i, \ x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{im} - x_{jm})^2} \tag{4}$$

where x_{im} denotes the m^{th} feature of i^{th} sample.

For a sample x_l in the test dataset, kNN runs throughout the whole dataset to compute the distance between x_l and each sample in the training dataset. The top k points that are closest to x_l will be recorded (e.g., named Set A) to classify x_l. Hence, the conditional probability of x_l belongs to class j can be estimated by Eq. (5).

$$P(y = jX = x_l) = \frac{1}{k}\sum_{i \in A} I(y_i = j) \tag{5}$$

where the k is the predefined k value; $I(y_i = j)$ equals 1 if instance y_i is in class j, otherwise it equals 0.

		Predicted results	
		Predicted condition **break**	Predicted condition **intact**
True condition	Condition **break**	True break (TB)	False intact (FI)
	Condition **intact**	False break (FB)	True intact (TI)

Fig. 1. Confusion matrix for pipe status classification.

2.5. Support vector classification (SVC)

SVC is a support vector machine for classification tasks by finding the optimal vector of hyperplane. The hyperplane is a plane that can divide the n-dimensional data points into two components. For instance, if the instances are a two-dimensional (2D) dataset, the hyperplane is a line on a 2D plane. SVC aims to find the hyperplane which could maximize the margins (sum of distances) from the hyperplane to the nearest training samples from each class [47]. Mathematically, SVC solves the following optimization Eqs. (6) to (8) to find the optimal hyperplane vector.

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C\sum_{i=1}^{n} \zeta_i \tag{6}$$

subject to

$$y_i\left(w^T \phi(x_i) + b\right) \geq 1 - \zeta_i \tag{7}$$

$$\zeta_i \geq 0, \ i = 1, \cdots, \ n \tag{8}$$

where the objective equation (Eq. (6)) is to maximize the margins (by minimizing $w^T w$) and minimize the misclassification of the distances (ζ_i) with a penalty term C. $\phi(x)$ is the kernel function that maps each sample into a higher dimensional space. w is the corresponding weight vector and b is the bias term. x_i is the training sample and y_i is the corresponding class.

With the trained weight vector (w) and bias term (b), the sample x in the test dataset can be classified by Eq. (9). SVC cannot output prediction probability directly, but such a probability (y) can be obtained by using the Platt scaling method [48].

$$y = sign\left(w^T \phi(x) w + b\right) \tag{9}$$

where sign is the signum function.

2.6. Metrics for ML model evaluation

The output of these five ML algorithms is a continuous value between 0 and 1, which denotes the confidence of a pipe being broken. Such confidence value is often interpreted as the failure probability [19]. Given that the ground truth of the test dataset is a binary class (intact or broken), a common approach to evaluate the prediction results is to divide them into each class based on a threshold, i.e., 0.5. In other words, the sample is predicted as break if the output is larger than 0.5, otherwise, it is 'intact'. Based on the predicted results and the ground truth, a confusion matrix can be obtained as shown in Fig. 1 [49], in which the terms 'True Break' and 'True Intact' represent the correctly classified samples. The 'False Intact' denotes the pipes that are predicted as intact but break. The 'False break' denotes the pipes that are predicted

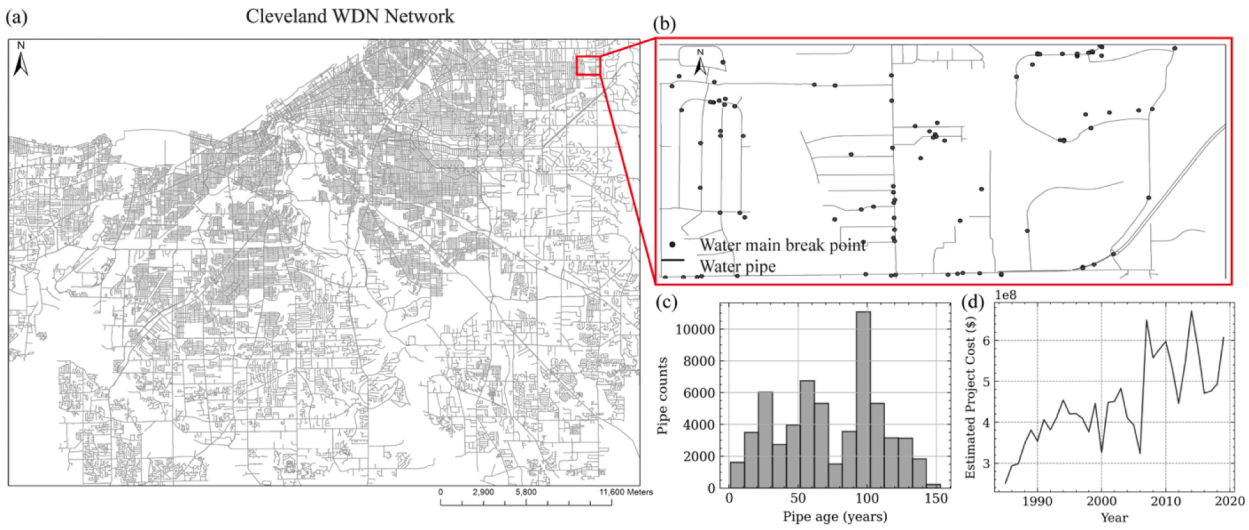


Fig. 2. Overview of the Cleveland WSN network (a: the pipes managed by Cleveland Water Department; b: example of maintenance records; c: distribution of pipes' service years; d: estimated annual maintenance cost ($)).

as break but intact.

Based on the categorized prediction results, three metrics are used in this study to quantitatively evaluate the performance of ML models, i.e., the accuracy, recall, and precision, as given in Eqs. (10)–(12).

$$Accuracy = \frac{TB + TI}{TB + TI + FB + FI} \tag{10}$$

$$Recall = \frac{TB}{TB + FI} \tag{11}$$

$$Precision = \frac{TB}{TB + FB} \tag{12}$$

here *TB* refers to True Break, TI refers to True Intact, FB refers to False Break, FI refers to False Intact.

The accuracy index measures the overall prediction accuracy by considering all the prediction results. The precision is the ratio between the True breaks and all the predicted breaks. The recall is the ratio between the True breaks and all real breaks. Specifically, a larger recall value means more break samples in the test dataset are successfully identified by the model. A larger precision value means more predicted break samples are true break samples. In reality, a low recall value may lead to the missing of the failure pipes and a low precision value may lead to mistakenly replacing intact pipes, thereby increasing unnecessary maintenance costs.

Previous studies indicated only using the Accuracy, Recall, and Precision is not enough for model evaluation [10]. To better compare the performance of different ML models, the Receiver Operating Characteristics (ROC) curve has been widely used in previous studies [10,20] since it describes the relationship between the FP rate and TP rate under different thresholds. However, the ROC curve could be misleading when it is applied in extremely imbalanced classification scenarios. The precision-recall curve (PRC) is suggested as the reliable alternative [50]. The PRC represents the relationship between recall and precision with different thresholds. To illustrate the different performance of ROC and PRC indexes with a highly imbalanced dataset, the Area Under Curve (AUC) values of ROC and PRC are calculated and compared in this study. The AUC value of each curve varies from 0.0 to 1.0, where 1.0 represents a perfect prediction, 0.5 indicates random guessing, and a value below 0.5 means worse than the random guessing.

3. Multi-Source data-aggregation framework for a large water supply network

Data preparation is an important step in data-driven ML approaches. The pipe information dataset and historical repairing records used in this study are collected by the Cleveland Water Department, which manages one of the largest WSNs in the United States. The collected dataset covers WSN that is used in Cuyahoga County which is the second-most populous county in the State of Ohio, US. The Cleveland Water Department is responsible for the water supply to 440,000 active user accounts with 5300 miles of water main pipelines. The studied area is also one of the largest cities in the Great Lakes area of North America. Due to the unique geology location, the soil of this area experiences frequent frozen and thawing processes during the winter days.

Fig. 2 shows the overview of the water supply system by the Cleveland Water Department. Fig. 2(a) shows the distribution of the pipe network, which includes a total number of 51,832 pipes. The physical attributes of each pipe, including the age, material, diameter, and length are included in the data record. Fig. 2(b) shows an example of pipe maintenance record in one of the areas in the WSN. The points are the locations where maintenance has been conducted. The maintenance date is also recorded by the Cleveland Water Department (CWD). Fig. 2 (c) shows the distribution of the pipe ages in the WSN. As can be seen, a large number of pipes have been used for over 100 years. Fig. 2(d) shows the estimated annual cost if all the damaged pipes were completed replaced. The estimation is based on the experience of the Cleveland Water Department, where the cost of renewing a pipe is estimated at around $483.74/feet. However, the real total cost may be lower than the estimated value as 1) multiple damages may be repaired at once, and 2) the maintenance might only fix the leak without completely replace the entire pipe. A method to provide reliable prediction of water pipe break will be very valuable for maintenance budget preparation.

To provide a comprehensive understanding of potential factors for WSN failures and associated maintenance, data from various sources are firstly aggregated. Although more and more data are becoming accessible in the public space nowadays, they are commonly hosted by different agents and stored in different formats, which makes it difficult to combine data from multiple sources. A noted gap in current implementation of data-driven approaches is the lack of guidelines on how to effectively assemble datasets from various sources. In this regard, this section is aimed at proposing a multi-source data-aggregation framework, as a supplement to the development of such guidelines. More specifically, Six datasets are considered in this study, i.e. (1) the WSN

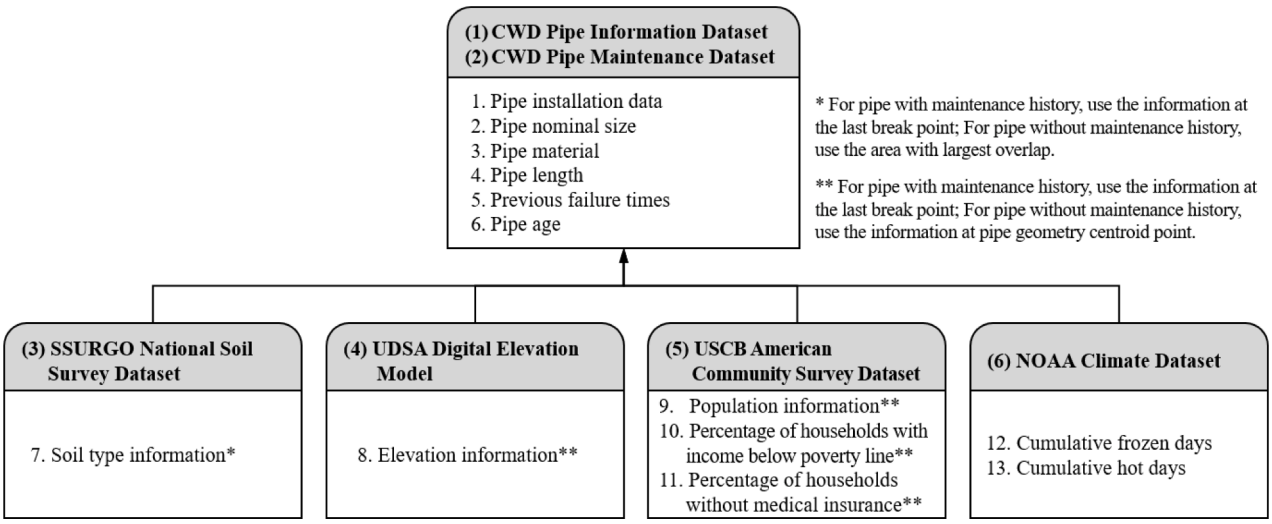


Fig. 3. Schema of aggregation of multiple sources of datasets.

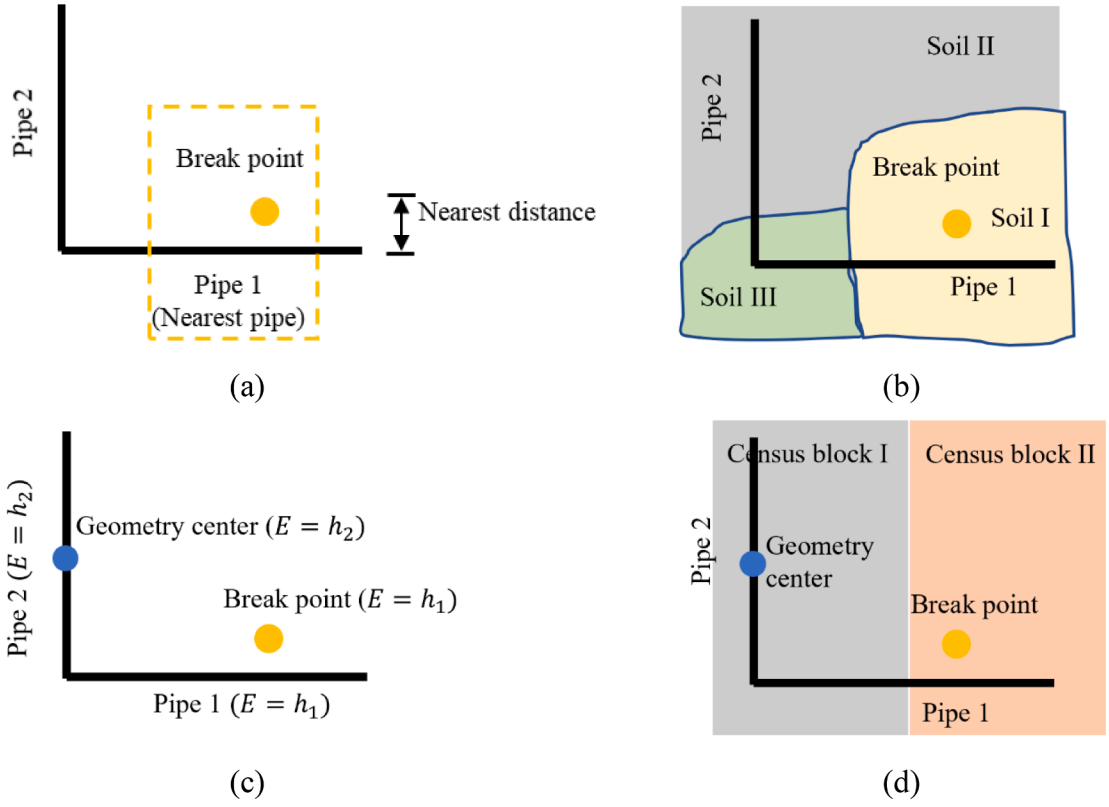


Fig. 4. Illustration of assumptions used for data aggregation process of different datasets: (a) pipe break dataset aggregated based on nearest distance of break point to a pipe, (b) soil type dataset aggregated based on geological information, (c) topographical dataset aggregation based on coordinate in the digital elevation model, and (d) census dataset aggregation based on physical coordinate of pipe or break point to the Census block.

pipe-information dataset, (2) historical pipe break-records dataset, both (1) and (2) are from the Cleveland Water Department (CWD), (3) the SSURGO soil type dataset from National Cooperative Soil Survey [51], (4) the topographical dataset from Digital Elevation Model (DEM) by United State Department of Agriculture (USDA) [52], (5) American Community Survey (ACS) 5-Year Data obtained from United States Census Bureau (USCB) [53], and (6) the climate dataset obtained from National Oceanic and Atmospheric Administration (NOAA) [54]. The latter four datasets are publicly accessible and are aggregated with the two datasets provided by the CWD to create the overall training and testing datasets. The data aggregation framework is outlined in Fig. 3. The criteria and assumptions for data aggregation are elaborated on below.

3.1. Pipe information and break records aggregation

Pipe information and break records are provided by CWD in the formats of two different layers in the Geographic Information Systems (GIS). As shown in Fig. 2(b), the pipe-information dataset is represented by 'lines' and the break-records dataset is represented by 'points'. Close

Table 1
Pipe physical factors.

Factor name	Description	Source
Pipe nominal size	North American set of standard sizes for pipes that consists of many varieties.	Pipe-information dataset
Pipe material	The materials of the pipe dataset.	
Pipe length	The length of the recorded pipes.	
Pipe age	The age of the pipe at the selected observation year.	Pipe-information dataset, and break-records dataset.
Previous break times	The break numbers before the selected observation year.	
Interval time to last break	Years between the observation year and the year of last break.	

examination indicated that these two types of data are not strictly overlapped. It is because the pipelines are based on actual laydown, while break locations (maintenance locations) are recorded manually or by GPS devices which do not align with the water pipes. For data aggregation, the break records are assumed to occur at its nearest pipe as long as the nearest distance is shorter than 1 m, which is the common resolution for GPS devices, as illustrated in Fig. 4(a). This allows the break records to be associated with the corresponding water pipes. After assigning the break records to the corresponding pipe, the physical factors about the pipe can be obtained for each pipe (i.e., Table 1). These include the pipe nominal size, pipe material, and pipe length that can be directly obtained from the original pipe information dataset. The pipe nominal size is a unitless designation standard size used in North America [55]. Pipes with any missing data of these three factors are removed. Another three factors including the pipe age, pipe previous break times, and interval time to last break are determined/processed based on the break records.

3.2. Soil type aggregation

As water pipes are commonly buried underground, previous studies have identified the important role of soil-pipe interaction on pipe failures [56–58]. In particular, the pipes may suffer failures resulting from the differential ground settlement [59]. The pipe deterioration process due to different soil corrosivity can also affect the pipe failure probability. Previous studies indicated the pipe corrosion occurs in a certain range of pondus hydrogen (pH) values [58]. As it is almost impossible to consider all the soil-related factors that may influence the pipe break, a practical way instead is to consider soil types that can be extracted from the public SSURGO dataset [51]. In this study, a total of 72 different soil types are considered. These soil types are classified based on several soil parameters including the slope, main components, surface texture, etc. [60]. The soil type data for individual pipes are assigned according to the major location of break records, together with the majority of the soil type where the pipe is embedded. For instance, in Fig. 4(b), the assigned soil type to Pipe 1 is Soil I since it is where the break point locates. For Pipe 2 that has no break records, the assigned soil type is Soil II since the majority part of Pipe 2 locates at Soil II.

3.3. Topographical data aggregation

As some of the pipe information data lack the operational water pressure, which may be an important factor for failure assessment of WSN, topographical data that describe the elevation of pipes can be considered to compensate for this lack because pipes located in higher elevations generally undergo lower water pressures. The area of the studied WSN pipes, Cuyahoga County, Cleveland, OH, experiences a diverse elevation range from 238 m to 401 m. The topographical dataset is obtained from Geospatial Data Gateway (GDG) of the USDA elevation dataset with a resolution of 30 m. The digital elevation model (DEM) provides the elevation of any single point in the Cuyahoga area. Similar to the soil type aggregation, pipe segments may experience different elevations, especially for the long-length pipes. Therefore, as illustrated in Fig. 4(c), the elevation of the breakpoint is used as the elevation of the pipe with break records (e.g., Pipe I). Otherwise, the elevation of the pipe geometry center is used.

Table 2
Considered census factors.

Factor name	Description	Source
Population	Estimated total population of the year 2019	ACS 5-year dataset (2019)
Poverty percentage	Percent of households whose income below the poverty line in the past 12 months	
Non health-insurance percentage	Percent of the population without any health insurance.	

3.4. Census data aggregation

As WSN is operated almost all the time for communities, it is reasonable to infer that community factors (such as user behaviors, population, etc.) may affect failure probabilities of pipes. To demonstrate this inference, public census data that contain community factors are adopted in this study. The census dataset obtained from the US census bureau contains enormous variables that describe every community block. The whole of Cuyahoga County is divided into 2952 community blocks in the census dataset. To consider the community factors from different perspectives, the population, poverty percentage, and non-medical insurance percentage, as listed in Table 2, are selected based on the availability of the dataset. Since a single pipe may cross multiple census blocks, a representative point defined in this study is used to assign the census data for each pipe. As illustrated in Fig. 4(d), for the pipe without break records, the representative point is chosen as the pipe geometry center (e.g., Census block I for Pipe 2), while for the pipe with break records, its breakpoint is used as the representative point (e.g., Census block for Pipe 1).

3.5. Climate data aggregation

Significant seasonal influence on the pipe's break has been observed in previous studies (e.g. air temperature, precipitation, etc.) [61, 62]. These studies indicate that the pipes suffered more breaks during extremely cold or hot days, possibly due to the soil-pipe interaction. However, previous studies did not account for the temperature data [20] for annual pipe break prediction. To overcome this limitation in considering the climate effects, the accumulated cold and hot days are used in this study to describe the climate experienced by each pipe during its service life. The hot days denote the days that have the highest temperature above 90 F while the cold days stand for those with the lowest temperature below 32 F (or the freezing index concept commonly used in cold region ground engineering). The number of these days are extracted from dataset provided by the climate service organizations such as NOAA. The cold and hot days experienced by each pipe used in this study are the accumulated days of each year from its installation date to the selected year of study (Eq. (13)).

$$Hot \Big/ Cold\ Days = \sum_{T=t}^{t+pipe\ age} D_T \tag{13}$$

where D_T is the total hot/cold days at year T, t is the pipe installation year.

The developed data aggregation framework assembles a comprehensive dataset that allows to integrate the physical properties of water pipes, the operational conditions, geology conditions, community socioeconomic characteristics, and climate conditions, which will be used

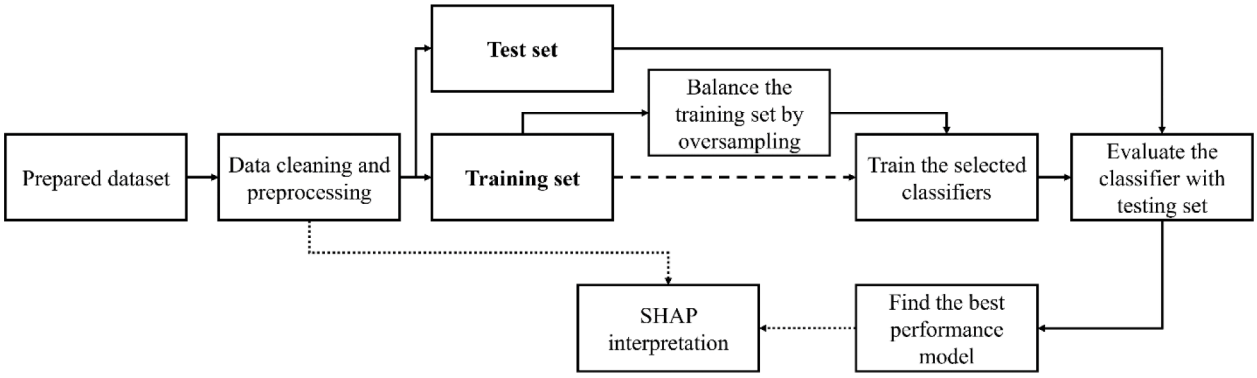


Fig. 5. Workflow for ML training and interpretation.

by the ML models to predict the pipe break behaviors.

4. Case study

4.1. Machine learning based pipe failure prediction workflow of machine learning modeling

Fig. 5 shows the overview of the developed workflow of the ML based water pipe break prediction. The aggregated dataset based on criteria discussed in Section 3 is selected and cleaned by removing the samples with any missing factors and outliers. To improve the performance of ML approaches, data normalization is applied to the numeric factors, which are standardized by removing the mean values and scaling to unit variance. The categorical factors will be encoded with one-hot encoding (a coding method to represent categorical factors by zeros and ones) if the ML algorithm cannot handle the categorical variables directly [63].

The prepared dataset was randomly split into training set and test set by the ratio of 8:2, which is a common approach to identifying the model's prediction ability. As the break records only account for around 10% of the total dataset, the issue of the imbalanced dataset for ML model training is an important issue to be taken care of. As shown in Fig. 5, both balanced and imbalanced training datasets are used for the model training. To balance the dataset, the oversampling method [64] is used in this study, which randomly replicates the minor class of pipe break dataset until the number of break samples equals the intact samples. The testing dataset is unchanged. Therefore, the models trained with both balanced and imbalanced datasets were evaluated with the same testing dataset. Once the ML model with the highest performance is identified, the whole dataset is used again for ML explanation by the use of an interpretation method named SHAP [65] to understand the influence of contributing factors on the pipe break. It should be noted the model interpretation results are independent of the train-test splitting method because the selected model is fitted again with the whole dataset. Details of data preparation and ML model performance evaluation are provided below.

4.2. Data preparation and assessment of data characteristics

4.2.1. Label assignment and dataset preparing

As the objective of this study is to predict the pipe status at a certain year, the pipe's status is defined either as break or intact, and therefore the ML model is defined as a classification problem. Preparing a dataset that covers each year of each pipe is unpractical and could lead to an extremely imbalanced dataset. To mitigate the level of imbalance as well as fully utilize the break records, the following procedures were applied to create the final dataset for ML training and testing.

1) For each pipe, a random year between its installation year and the latest update before this study (Oct 2020) is selected. Time-dependent factors during this period (i.e., previous failure times, pipe age, cumulative frozen and hot days experienced) are determined.
2) The pipe status recorded at the selected year is labeled as 0 for intact status or 1 for break status.
3) To fully utilize the pipe break records, for any pipe with break record (s), the data of its last break year is added to the dataset created in the previous steps.
4) In the final prepared dataset, the prepared dataset is split into training and test dataset with a ratio of 8:2.

Table 3
Categories and statistical properties of factors considered for pipe break prediction.

Factor	Type of factors	Units	Mean	Std.	Min	Max
Nominal size	Physical	–	8.3	2.1	2	12
Material		–	Ductile Iron, Cast Iron, and Unknown			
Length		feet	326.9	322.3	0.5	1332.5
Age		years	44	29.77	1	127
Previous break times	Operational	–	0.4	1.6	0	37
Years to last break time		years	1.2	3.9	0	34
Soil type	Environmental	–	UmB, MmB, UeA, Ub, EsC, …, etc. 72 different soil types			
Elevation		m	259.3	46.4	174.8	382.9
Cold days		days	1861.6	1236.7	33	5378
Hot days		days	462.2	318.8	0	1219
Population	Societal	–	3432.5	1294.0	703	7953
Poverty percentage		%	12.6	13.3	0.4	91.1
Percent without health insurance		%	5.0	3.4	0.2	18

The selected data are subjected to the data aggregation process described in the earlier part of this paper. Data cleaning is then conducted, where data points with missing information are removed. After data cleaning, 39,491 pipe data samples are obtained including 31,078 non-break samples and 8413 break samples.

In light that the random selection of a year in Step 1) may influence the model prediction results. Therefore, ten random selections are conducted, and the average values of evaluation matrices are used for model performance assessment.

4.2.2. Characteristics of factors in aggregated dataset

To present an overview of the behaviors of aggregated data, Table 3 summarizes an example of their categories and statistical properties including mean, standard deviation (std.), minimum, and maximum

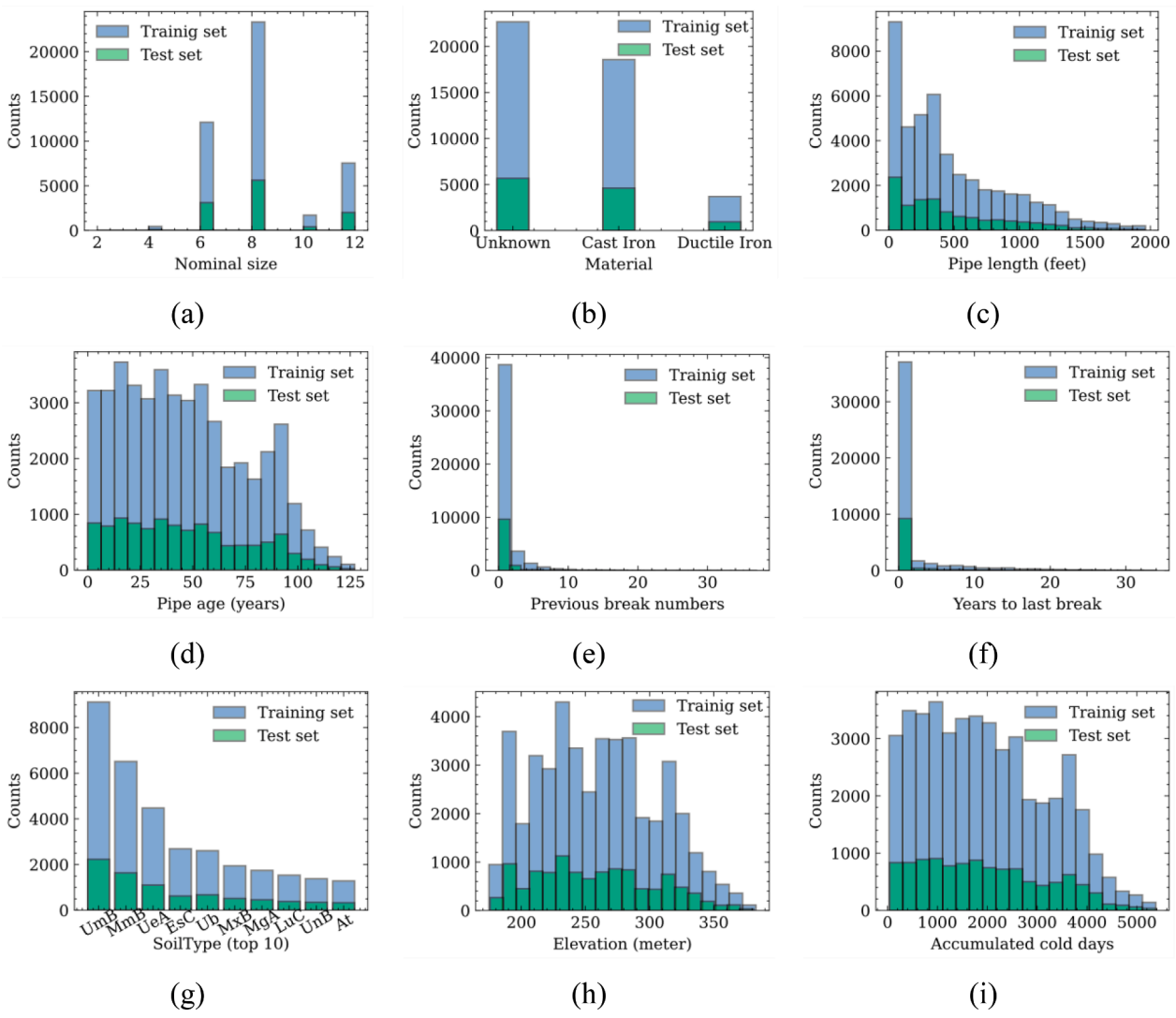


Fig. 6. Histograms of considered factors in terms of training and test sets.

from one dataset. A total of 13 factors are aggregated including 11 continuous variables and two categorical variables. These factors are grouped into four main types, i.e., physical, operational, environmental, and societal.

Fig. 6 shows histograms of the considered factors in both training and testing datasets. In general, features of the testing dataset are well covered by the training dataset. More specifically, distributions of the physical variables are shown in Fig. 6(a) to (d). Note that Ductile Iron and Cast Iron are the most widely used materials in WSN systems, while there are also a large number of pipes whose materials are unknown (Fig. 6(c)). Fig. 6(e) and (f) illustrate the characteristics of two operational variables. Fig. 6(g) to (j) show the four environmental variables. 72 different soil types are assigned to the pipes, but for the convenience of visualization, Fig. 6(g) only plots the 10 most popular ones, i.e., UmB (Urban land-Mahoning complex), MmB (Mahoning-Urban land complex), UeA (Urban land-Elnora complex), Ub (Urban land), EsC (Ellsworth-Urban land complex), MxB (Mitiwange-Urban land complex), LuC (Loudonville-Urban land complex), UnB ((Urban land-Mitiwanga complex), UoB (Urban land-Oshtemo complex), and MgA (Mahoning silt loam). The elevation of the pipes varies from 150 to 382 m (Fig. 6 (h)). The considered climate factors include the accumulated hot days and cold days as shown in Fig. 6(i) and Fig. 6(j). It is interesting to find that they share similar trends with the distribution of pipe age, which indicates that the climate factors are critical for pipe failures. As shown in Fig. 6(k) to (m), the population, percentage of households below poverty lines, and percentage of households without health insurance are chosen to represent socio-economical features of each pipe for the served community. It is worth noting that the population within each Census community block range from 700 to around 8000, since Cuyahoga County includes some densely populated area such as Cleveland City. For most community blocks, the poverty percentage is below 40%. However, the most poverty block has more than 90% households whose annual incomes are below the poverty line. Similar to observed poverty condition, the percent of people in community blocks who do not have any medical insurance ranges from 0.2% to 18%.

Based on whether this pipe is broken in the selected observation year, the group of pipes has been divided into two classes, the class code is 1 if the pipe breaks in the observation year and the code is 0 if it is intact. The results are shown in Fig. 6. (n).

4.2.3. *Relationship between internal factors and break status*

The correlation matrix can be used to show the internal relationship among the factors considered to contribute to pipe failures as well as the external relationship between each factor and the target. Since the datasets contain both numerical and categorical variables, different correlation indicators are used to quantify their correlation. Methods to

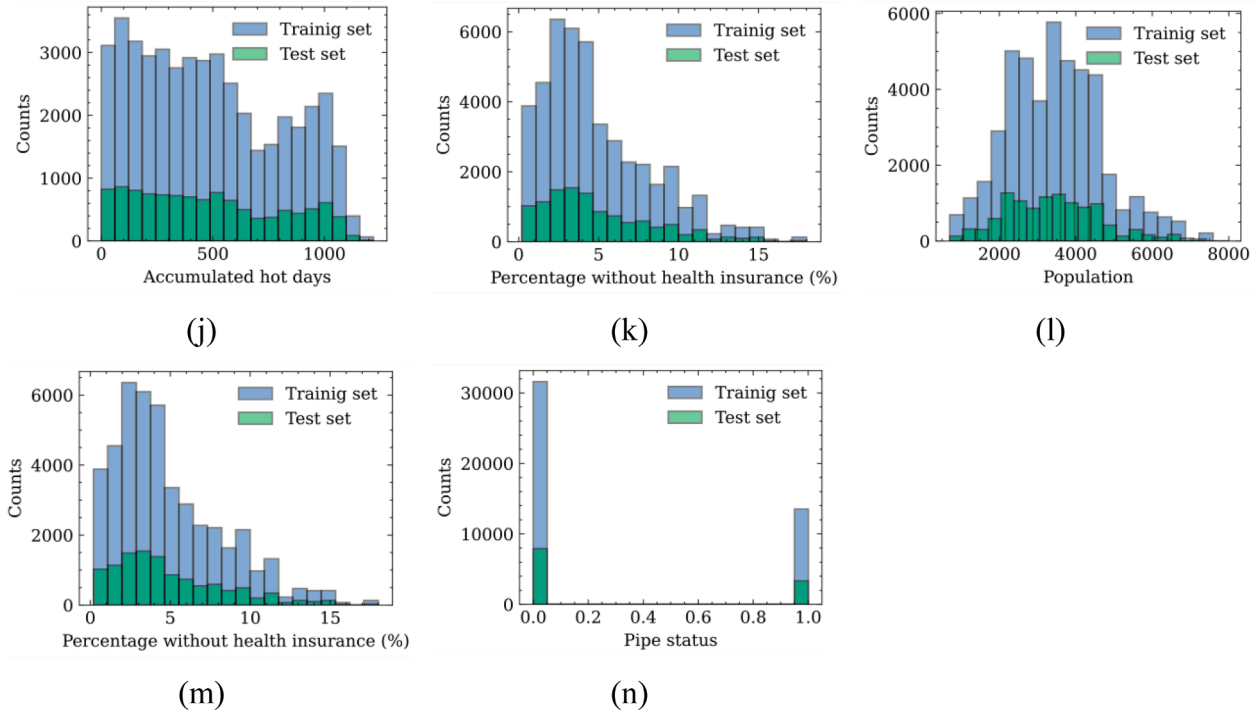


Fig. 6. (*continued*).

	Continuous variable	**Categorical variable**
Continuous variable	Pearson correlation	Correlation ratio
Categorical variable	Correlation ratio	Cramér's V

Fig. 7. The correlation indicators used for numerical variables and categorical variables.

determine the correlations between different types of variables are shown in Fig. 7. Specifically, the correlation between two numerical variables is represented by Pearson correlation [66], while that between two categorical variables is quantified by the Cramér's V coefficient [67]. Besides, the Correlation ratio is used between the categorical and numerical variables [68]. These indicators fall between −1 to 1 for numerical factors and 0 to 1 when one or more factors are categorical (1 or −1 denotes complete positive or negative association while 0 denotes no association). Detailed discussion about the correlation computations is not shown here due to the paper length limitation. Interested readers could refer to the Supplementary file (Algorithm I).

Fig. 8 shows the final correlation matrix among the 13 factors considered as well as their correlation with the pipe break status (indicated as 'target' in the figure). Factors with the highest correlation values to the target (pipe break status) are the hot days, cold days, and pipe ages. These imply that the weather and pipe service age are among the most important factors determining pipe conditions. The other factors that follow include the previous break number, interval year, and pipe length, etc. The present study considers all of them to observe their behaviors in the ML model and model interpretation. The correlation map in Fig. 8 also indicates that some of the considered societal variables (population, percentage of poverty, percentage without health insurance) are correlated but not strongly, which is good for them to be treated as independent variables in the ML model. Besides, there is no correlation between any single input variable and the output variable that is dominant, which indicates that the pipe failure is a complex problem that does not depend on any single factor.

4.3. L-based pipe failure prediction

4.3.1. Prediction results

The five types of supervised ML algorithms for classification problems as introduced in the Background section are developed for the pipe break classification problem. The hyperparameters of each ML algorithm are optimized by grid-search optimization [69]. For the ANN model, one input layer, a dropout layer, two hidden neural network layers, and an output layer are used. The neuron numbers of the hidden layers are 64 and 128 respectively. Fig. 9a) shows the prediction results when the model is trained with an imbalanced dataset and Fig. 9b) shows the results when the model is trained with balanced dataset. The recall and precision matrices for each class are denoted at the right and bottom sides, respectively. The overall model accuracy is denoted at the right-bottom cell. Regardless of balanced or imbalanced dataset, the LightGBM and ANN models achieve the best prediction in terms of the accuracy, recall, and precision metrics. On the contrary, the KNN and SVC models tend to miss many break samples when trained with the imbalanced dataset and miss many intact samples when trained with the balanced dataset. Oversampling method helps the model identified more break samples.

To illustrate the overall performance of different models, the average Receiver Operating Characteristics (ROC) and precision-recall curve (PRC) metrics and the corresponding area under curve (AUC), and average training time of ten sampled datasets are calculated and presented in Fig. 10. Among all the considered models, the LightGBM model achieved the highest ROC and PRC values and a short training time regardless of imbalanced or balanced training sets. A potential reason is the LightGBM can deal with categorical variables without one-hot-

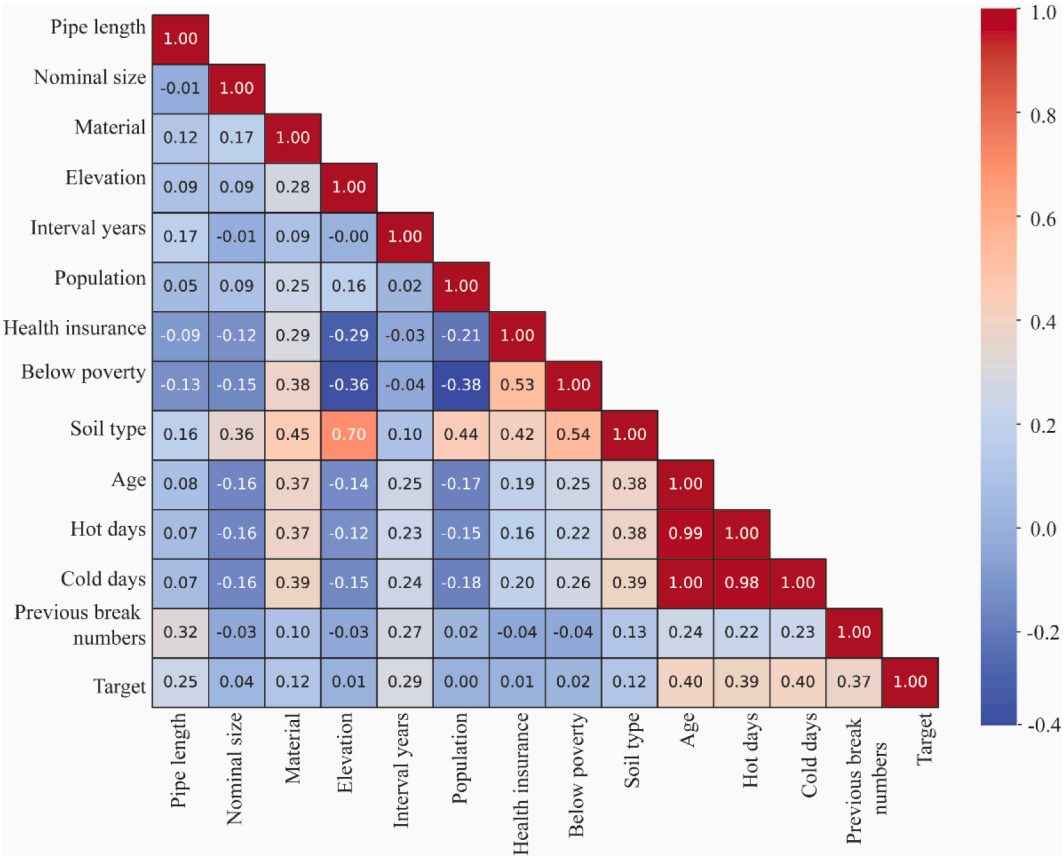


Fig. 8. Correlation map of considered factors and prediction target.

encoding so the input dataset is much less sparse than that of other models. When comparing the ROC matrix and PRC matrix, it can be observed that the LightGBM, ANN, LR, and kNN achieved similar ROC values, which is controversial to the previously observed results (see Fig. 9). The PRC matrix shows more obvious different values for different models, which demonstrates a better identification of the optimal models. Based on the PRC results, the balanced training LightGBM model achieved a higher AUC of PRC value (0.810) than that of imbalanced training model (0.793).

Figs 8, 9, and 10 represent the results when using randomly based train-test splitting. However, one critical question in engineering application is that whether the machine learning model trained by data recorded from previous events could effectively predict the future events. To address this concern, a time-line based train-test splitting method is used again to compute the above processes. The results are not discussed here due to the similar conclusions. Interested readers could refer Fig. S1 to Fig. S3 in the supplementary material.

To better understand the performance of using different ML models for pipe break prediction, the performance of these ML models is compared under different indicators in Table 4 based on the accuracy, training speed, ability to handle the categorical features, and inherent model interpretability. Based on the classification results with imbalanced and balanced training datasets, it can be confirmed that the LightGBM achieved the best accuracy on the current dataset. The areas under PRC for the test dataset are about 0.82 when trained with the balanced dataset. Then followed by the ANN, LR, SVC, and kNN models. In terms of the computational efficiency, the LR model finished the learning process within a few seconds, then followed by the LightGBM model and ANN model. The kNN model and SVC model took the longest time among these models. The LightGBM model is the only model that could analyze the categorical variables without any encoding due to its histogram and leaf-wise tree growth strategy feature. Although the inherent explanation ability of different models is not the focus of this study, they are discussed here based on the previous studies [20,42] for completeness. Generally, the statistical algorithms (e.g., the LR model) are considered the most explainable algorithms since the weights of each factor are transparent and can be interpreted based on their physical meaning. The Logistic-based algorithms such as LightGBM model are also quite easy to interpret as the relative importance of factors can be quantified based on their roles during the tree splitting. Finally, the perceptron-based algorithms (e.g., ANN model), instance-based learning algorithms (e.g., kNN), and support vector machines (i.e., SVC) are considered as black box model with poor interpretability.

4.3.2. *Interpretation of considered factors with shap method*

Based on the performance comparison of the five major types of ML models for classification problems as described in the previous sections, the LightGBM model features both the highest prediction accuracy and is also computationally efficient. Although the model itself has a moderate explanation ability, it is still unable to fully understand the contribution of each factor to the output directly. To solve this issue, an ML model interpreter, Shapley Additive exPlanations (SHAP) [65] is adopted together with the LightGBM model trained with the balanced training dataset, to interpret the contribution of each factor. SHAP presents a way to calculate the additive feature importance score for each factor [70]. The higher the importance score, the more important the factor to the final ML model prediction. The SHAP interpretation method together with decision-tree based ML algorithms have been widely used in the field of civil engineering, including some scenarios where highly correlated variables existed, such as the explanation of the failure of reinforced concrete [71], the explanation of RC walls shear strength [72], and the roadway segment crashes [73].

The impacts of each factor on the pipe break can be gathered to evaluate their overall influence. Fig. 11 shows the overall importance of

		Predicted results		
		I	B	Recall
True condition	I	6023	489	0.925
	B	694	1293	0.651
Precision		0.897	0.726	0.861

LightGBM model

		Predicted results		
		I	B	Recall
True condition	I	5681	831	0.872
	B	585	1402	0.705
Precision		0.906	0.628	0.833

ANN model

		Predicted results		
		I	B	Recall
True condition	I	5032	1480	0.773
	B	381	1606	0.808
Precision		0.929	0.520	0.781

LR model

		Predicted results		
		I	B	Recall
True condition	I	5992	520	0.920
	B	934	1053	0.530
Precision		0.865	0.669	0.828

kNN model

		Predicted results		
		I	B	Recall
True condition	I	5975	537	0.918
	B	890	1097	0.552
Precision		0.870	0.671	0.832

SVC model

a) Test set prediction results by model trained with imbalanced data

		Predicted results		
		I	B	Recall
True condition	I	5390	1082	0.833
	B	473	1591	0.771
Precision		0.919	0.595	0.818

LightGBM model

		Predicted results		
		I	B	Recall
True condition	I	4849	1623	0.749
	B	332	1732	0.839
Precision		0.935	0.516	0.771

ANN model

		Predicted results		
		I	B	Recall
True condition	I	5027	1445	0.776
	B	458	1606	0.778
Precision		0.916	0.526	0.777

LR model

		Predicted results		
		I	B	Recall
True condition	I	4803	1669	0.742
	B	399	1665	0.807
Precision		0.923	0.500	0.757

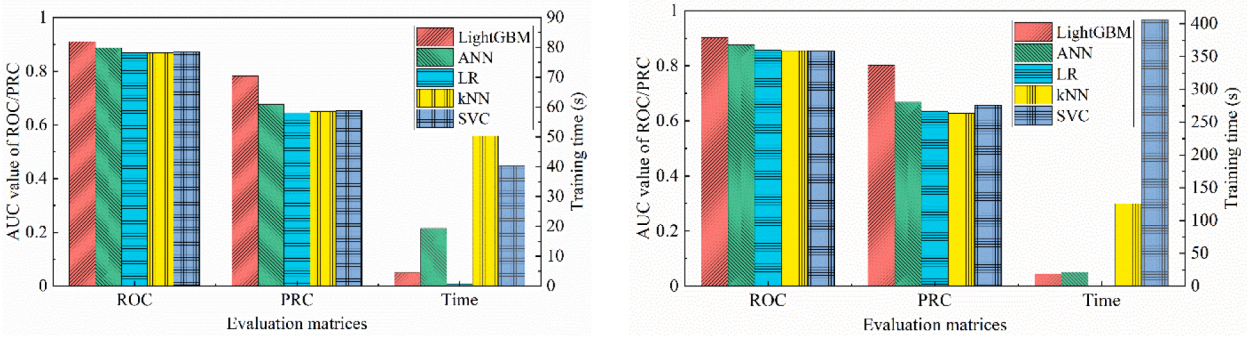
kNN model

		Predicted results		
		I	B	Recall
True condition	I	4994	1478	0.772
	B	447	1617	0.783
Precision		0.917	0.522	0.774

SVC model

b) Test set prediction results by model trained with balanced data

Fig. 9. Prediction results with imbalanced and balanced training dataset ('I' shorts for intact and 'B' shorts for break).



a) trained with imbalanced dataset b) trained with balanced dataset

Fig. 10. Comparison of applied ML models using different evaluation metrics.

the considered factors. The Pipe age, Cold days, and Hot days are kept in the dataset to study the impact of climate factors. The result shows the 'Cold days' has the most impact among them, followed by the 'Hot days' and 'Pipe age'. The interval to the last break, Cold days, and pipe length turns out to be the most influential factors. The pipe material and the nominal size of the pipe have a relatively lower influence on the pipe

Table 4
Model summarization for pipe break prediction.

	LightGBM	ANN	LR	kNN	SVC
Accuracy in general	***	**	*	*	*
Speed of learning process	**	**	***	*	*
Handle categorical variables	***	*	*	*	*
Inherent model interpretability	**	*	***	*	*

***denote the best performance.
*denote the worst performance.

break probability.

The following figures show the detailed effects of each factor on the pipe break probability. The SHAP method computes an impact value for each variable at each sample. To represent the overall impact of considered factors on the pipe break probability, each factor's impact values are extracted from all pipe samples. The impact of continuous variables is colorized for their magnitudes. The impact variables are represented by their mean values. The following conclusions can be inferred based on the observations. Many of these conclusions are consistent with previous studies, which demonstrate the soundness of the model interpretation results.

1) Fig. 12 shows the impact of physical factors considered. All these factors have a positive influence on the pipe failure probability, i.e., the larger the value of the factor, the larger the SHAP value. The distribution of pipe length values indicates the longer pipes have a higher failure probability than shorter ones. Similar observation can be found on pipe age. Older pipes correspond to larger SHAP value, indicating they are more prone to failure than the younger ones. The nominal size also shows a positive correlation to the SHAP value, which means that pipes with larger nominal sizes have higher failure probabilities. The impact of pipe material is shown separately on the right side of Fig. 12 since it is a categorical variable. Although both with small SHAP values, the Cast Iron corresponds to positive SHAP impact value while Ductile Iron corresponds to negative SHAP impact value. This indicates that pipes made with Cast Iron correspond to higher break probability compared with those made of Ductile Iron. This is partially attributed to the fact that Ductile Iron is less brittle than Cast Iron.

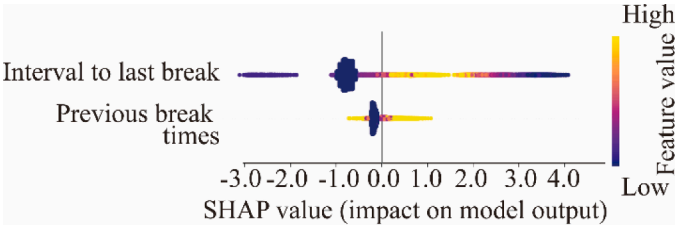


Fig. 13. Plot of SHAP values showing the impacts of operational factors.

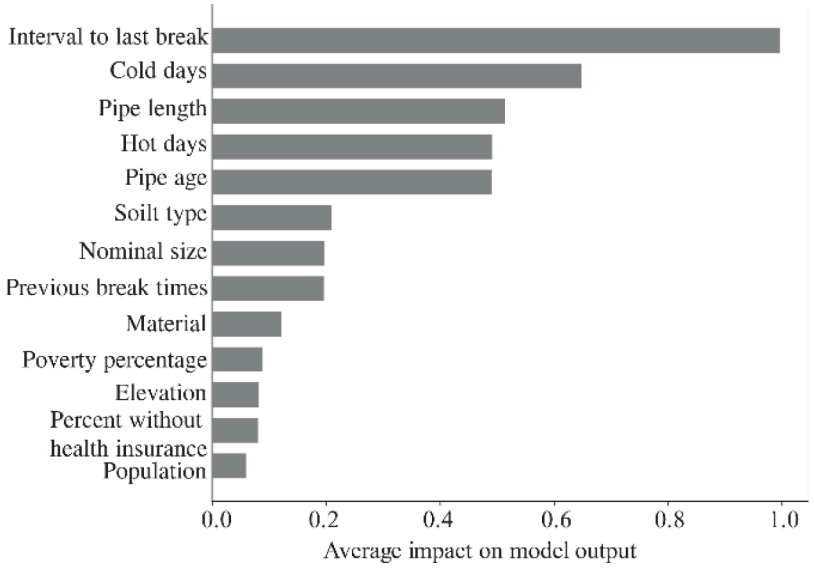


Fig. 11. The overall rank of considered factors on pipe failure probability.

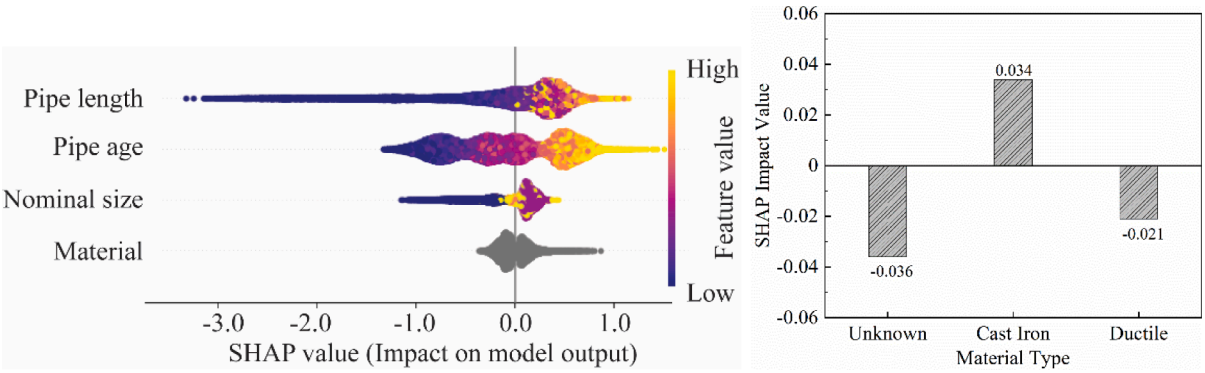


Fig. 12. The impacts of physical factors.

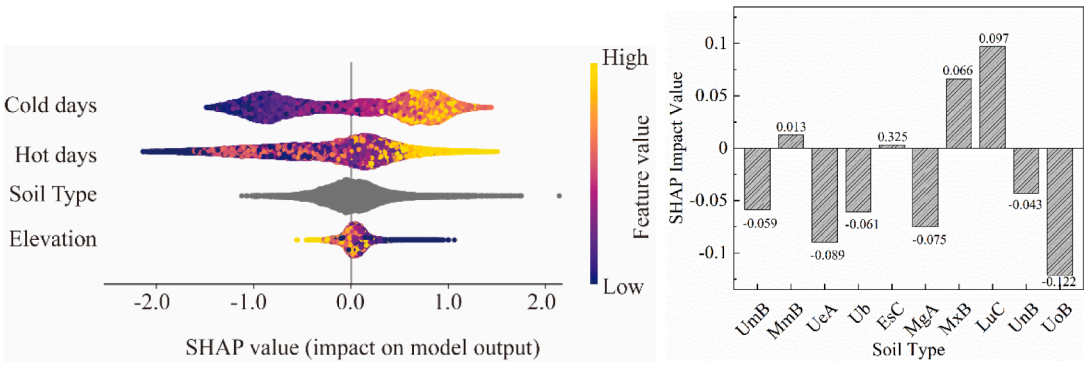


Fig. 14. Environmental factors impact.

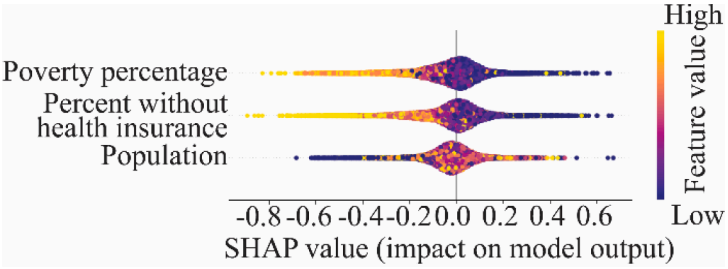


Fig. 15. Community factors impact.

2) Fig. 13 shows the effects of service operational factors. Compared with the times of previous breaks, the time interval since the last break shows higher impact on the pipe failure probability. While a larger time interval since the last break on the pipe is found to corresponded to increased failure probability, a smaller time interval shows both negative and positive influences (rather than only negative influence as intuitively) on the failure probability. This implies that there are quite a few pipes broken soon after the installation or repair. This observation fits previous studies that indicated the failure rate of a pipe is a bathtub curve, i.e., high failure probabilities occur either at the early life or at wear-out life of pipes [74]. Besides, the results of SHAP values in Fig. 13 show that a larger number of previous breaks at the pipe corresponds to a higher failure probability at the pipe. This is consistent with empirical observations by interviewing with practitioners, i.e., pipe failures tend to occur more frequently at similar locations.
3) For the environmental factors shown in Fig. 14, the results demonstrate that the larger the number of cold days a pipe experienced, the higher the probability it broke. It may be induced by the influence of soil settlement due to the soil freezing and thawing process. The impact of soil types is shown on the right side of Fig. 14. Among the top 10 soil types that most pipes are buried, the LuC (Loudonville-Urban land complex) soil triggers the highest pipe failure probability while the UoB (Urban land-Oshtemo complex) shows the lowest failure probability. Finally, the result indicates that the pipes in higher elevations are less likely to break, which is reasonable because pipes located at higher elevations experience lower service water pressure. This leads to less internal stress on the pipe by the service water pressure.
4) Regarding the societal factors as shown in Fig. 15, it is surprising to find that communities that are poorer or with less health insurance coverage have a smaller probability of water pipe break. It may be because the poverty areas are less inspected, or they generally use less frequency/amount of water that helped to extend the service life of the pipe. In-depth reasons require more studies based on more data. Moreover, the density of a community block does not have a consistent trend of effects on the pipe failure probability, although the densely populated area corresponds to a relatively higher water pipe break probability in most pipe samples.

5. Conclusions

This study aims to explore the ML techniques to predict water pipe breaks based on data from a large WSN and to understand the effects of contributing factors. A framework that integrates WSN maintenance datasets and multiple public datasets is proposed, which is a critical step that allows considering contributing factors for pipe failures, including the geology, climate, and socioeconomic factors. With the aggregated data, five different ML models, each from one of the five major types of ML classification models, are developed for pipe break prediction. The models are trained with an imbalanced dataset where the majority are from intact pipes, as well as a balanced training dataset, where the oversampling method is used to balance the training dataset. The results of different training datasets are compared. Finally, the trained ML model is interpreted by an interpolation method, SHAP. Major contributions of this study include:

1) It provides a state-of-the-art data aggregation framework that integrates multi-source public datasets, leading to the largest real-field dataset (both in size and timeline) and associated largest number of input parameters for machine learning modeling for WSN. Hence, this study significantly expands and deepens the communities' understanding on the effects of engineering, geology, climate and socioeconomic factors. To our best knowledge, this is the first paper to assess the influence of socioeconomic factors on the failure of water supply networks.
2) For the sake of implementation, five popular machine learning algorithms are examined and comprehensively compared by five metrics (i.e., the accuracy, computation time, influence of categorical variables, and interpretation ability). The LightGBM model achieved the highest performance with the second shortest training time. Meanwhile, the Receiver operating characteristics (ROC) is demonstrated to be too optimistic about the results than the

precision-recall curve (PRC) metric when the dataset is highly imbalanced.

3) The SHAP interpretation results fit with previous studies, which demonstrated its ability to interpret the influence of the contributing factors. The results indicate the significant influence of pipe buried time, especially the interval time to the last break, the experienced cold days, hot days, and pipe age. The contribution of pipe physical factors and climate factors are in accordance with results reported in previous studies. The contribution of community characteristics indicates that areas with high poverty are associated with a lower pipe break probability (or less frequently maintained), while areas with high population density correspond to a higher probability of water pipe break. These indicate that socioeconomic factors have an important influence on the pipe service conditions.

Finally, water pipe failure is a result of the complex nonlinear interactions among various factors. Previous studies often simplified the analysis process by considering a small number of factors due to the model capability and data availability. Future studies should consider to include more advanced ML techniques and more extensive dataset to further improve the reliability of pipe failure prediction. Progresses in these areas will potentially catalyze the transformation of decision-making support for proactive management of WSN to achieve sustainability goal.

Declaration of Competing Interest

None.

Acknowledgment

The authors acknowledge the help provided by the Cleveland Water Department during the course of this study. The assistance of Mr. Alex Margevicius, the commissioner of CWD, is highly appreciated. The study is partially supported by the US National Science Foundation (No. 1638320).

Author Statement

Machine Learning based Water Pipe Failure Prediction: The Effects of Engineering, Geology, Climate and Socio-Economic Factors.

Xudong Fan: model, data collection and integration, results analyses, draft manuscript. Xijin Zhang: assist with data analyses. Xiaowei Wang: proofreading. Xiong (Bill) Yu: Envision of the study, methodology, research tactics, editing and proofread manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ress.2021.108185.

Reference

[1] Folkman, S., Water main break rates in the USA and Canada: a comprehensive study. 2018.
[2] Najjaran, H., R. Sadiq, and B. Rajani, Modeling pipe deterioration using soil properties-an application of fuzzy logic expert system, in Pipeline Engineering and Construction: what's on the Horizon? 2004. p. 1–10.
[3] Association AWW. Dawn of the replacement era: reinvesting in drinking water infrastructure: an analysis of twenty utilities' needs for repair and replacement of drinking water infrastructure. American Water Works Association; 2001. May 2001.
[4] Kim M-Y. Residual lifetime assessment of cold-reheater pipe in coal-fired power plant through accelerated degradation test. Reliab Eng Syst Saf 2019;188:330–5.
[5] Aryai V. Time-dependent finite element reliability assessment of cast-iron water pipes subjected to spatio-temporal correlated corrosion process. Reliab Eng Syst Saf 2020;197:106802.
[6] Barton NA. Improving pipe failure predictions: factors affecting pipe failure in drinking water networks. Water Res 2019;164:114926.
[7] Kettler A, Goulter I. An analysis of pipe breakage in urban water distribution networks. Can J Civ Eng 1985;12(2):286–93.
[8] Yamijala S, Guikema SD, Brumbelow K. Statistical models for the analysis of water distribution system pipe break data. Reliab Eng Syst Saf 2009;94(2):282–93.
[9] Pietrucha-Urbanik K. Failure analysis and assessment on the exemplary water supply network. Eng Fail Anal 2015;57:137–42.
[10] Debón A. Comparing risk of failure models in water supply networks using ROC curves. Reliab Eng Syst Saf 2010;95(1):43–8.
[11] Kabir G, Tesfamariam S, Sadiq R. Predicting water main failures using Bayesian model averaging and survival modelling approach. Reliab Eng Syst Saf 2015;142:498–514.
[12] Shirzad A, Tabesh M, Farmani R. A comparison between performance of support vector regression and artificial neural network in prediction of pipe burst rate in water distribution networks. KSCE J Civ Eng 2014;18(4):941–8.
[13] Jafar R, Shahrour I, Juran I. Application of artificial neural networks (ANN) to model the failure of urban water mains. Math Comput Model 2010;51(9–10):1170–80.
[14] Ray PA. Multidimensional stress test for hydropower investments facing climate, geophysical and financial uncertainty. Global Environ Change 2018;48:168–81.
[15] Wang X. Empirical probability distribution models for soil-layer thicknesses of liquefiable ground. J Geotech Geoenviron Eng 2021;147(6):06021005.
[16] Laucelli D. Study on relationships between climate-related covariates and pipe bursts using evolutionary-based modelling. J Hydroinf 2014;16(4):743–57.
[17] Rajani B, Tesfamariam S. Uncoupled axial, flexural, and circumferential pipe soil interaction analyses of partially supported jointed water mains. Canadian geotechnical journal 2004;41(6):997–1010.
[18] Guidotti R, Gardoni P, Rosenheim N. Integration of physical infrastructure and social systems in communities' reliability and resilience analysis. Reliab Eng Syst Saf 2019;185:476–92.
[19] Winkler D. Pipe failure modelling for water distribution networks using boosted decision trees. Struct Infrastruct Eng 2018;14(10):1402–11.
[20] Robles-Velasco A. Prediction of pipe failures in water supply networks using logistic regression and support vector classification. Reliab Eng Syst Saf 2020;196:106754.
[21] Committee AB. ASME B31G-2009: manual for determining the remaining strength of corroded pipelines. Am Soc Mech Eng 2009.
[22] Netto T, Ferraz U, Estefen S. The effect of corrosion defects on the burst pressure of pipelines. J Constr Steel Res 2005;61(8):1185–204.
[23] Wang N, Zarghamee MS. Evaluating fitness-for-service of corroded metal pipelines: structural reliability bases. J Pipeline Syst Eng Pract 2014;5(1):04013012.
[24] Cronin DS, Pick RJ. Experimental database for corroded pipe: evaluation of RSTRENG and B31G. In: in 2000 3rd International Pipeline Conference. American Society of Mechanical Engineers Digital Collection; 2000.
[25] Rajani B, Kleiner Y. Comprehensive review of structural deterioration of water mains: physically based models. Urban Water 2001;3(3):151–64.
[26] ASME B31 Committee. ASME B31G-2009: manual for determining the remaining strength of corroded pipelines. Am Soc Mech Eng 2009.
[27] Mazumder RK, Salman AM, Li Y. Failure risk analysis of pipelines using data-driven machine learning algorithms. Struct Saf 2021:89.
[28] Kleiner Y, Rajani B. Comprehensive review of structural deterioration of water mains: statistical models. Urban water 2001;3(3):131–50.
[29] Martins A, Leitão JP, Amado C. Comparative study of three stochastic models for prediction of pipe failures in water supply systems. J Infrastruct Syst 2013;19(4):442–50.
[30] Osman H, Bainbridge K. Comparison of statistical deterioration models for water distribution networks. J Perform Constr Facil 2011;25(3):259–66.
[31] Scheidegger A, Leitao JP, Scholten L. Statistical failure models for water distribution pipes–A review from a unified perspective. Water Res. 2015;83:237–47.
[32] Constantine A, Darroch J. Pipeline reliability: stochastic models in engineering technology and management. World Scientific Publishing Co; 1993.
[33] Scheidegger A. Extension of pipe failure models to consider the absence of data from replaced pipes. Water Res 2013;47(11):3696–705.
[34] Tang K, Parsons DJ, Jude S. Comparison of automatic and guided learning for Bayesian networks to analyse pipe failures in the water distribution system. Reliab Eng Syst Saf 2019;186:24–36.
[35] Zhang Y, Weng W. Bayesian network model for buried gas pipeline failure analysis caused by corrosion and external interference. Reliab Eng Syst Saf 2020;203:107089.
[36] Ren C-y, Qiao W, Tian X. Natural gas pipeline corrosion rate prediction model based on bp neural network, in fuzzy engineering and operations research. Springer; 2012. p. 449–55.
[37] Peng X-y, Zhang P, Chen L-q. Long-distance oil/gas pipeline failure rate prediction based on fuzzy neural network model. Computer Science and Information Engineering 2009. 2009 WRI World Congress onIEEE.
[38] Tabesh M. Assessing pipe failure rate and mechanical reliability of water distribution networks using data-driven modeling. J Hydroinf 2009;11(1):1–17.
[39] Christodoulou S, Deligianni A. A neurofuzzy decision framework for the management of water distribution networks. Water Resour Manage 2010;24(1):139–56.
[40] Chen TY-J, Guikema SD. Prediction of water main failures with the spatial clustering of breaks. Reliab Eng Syst Saf 2020;203:107108.
[41] Jara-Arriagada C, Stoianov I. Pipe breaks and estimating the impact of pressure control in water supply networks. Reliab Eng Syst Saf 2021;210:107525.
[42] Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: a review of classification techniques. Emerg Artif Intellig Appl Comput Eng 2007;160(1):3–24.

[43] Ke G. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30:3146–54.
[44] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943;5(4):115–33.
[45] Gasso G. Logistic regression. INSA Rouen-ASI Departement Laboratory; 2019.
[46] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967;13(1):21–7.
[47] Maldonado S. Feature selection for support vector machines via mixed integer linear programming. Inf Sci (Ny) 2014;279:163–75.
[48] Platt J. Probabilistic outputs for SVMs and comparisons to regularized likelihood methods. Adv Large Margin Classifiers 2004:61–74.
[49] Stehman SV. Selecting and interpreting measures of thematic classification accuracy. Remote Sens Environ 1997;62(1):77–89.
[50] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 2015;10 (3):e0118432.
[51] NRCS, Soil survey staff, natural resources conservation service, United States department of agriculture. Soil Survey Geographic (SSURGO) Database for northeast Tennessee., 2010.
[52] Abdollahian, Data release for results of societal exposure to California's volcanic hazards (ver. 3.0, November 2019): U.S. Geological Survey data release, accessed February 10, 2020. https://doi.org/10.5066/F7W66JRH., 2018.
[53] U.S. Census Bureau, American Community Survey 5-Year Data (2009-2019). Retrieved from https://www.census.gov/data/developers/data-sets/acs-5year.html, 2020.
[54] Vose RS. Improved historical temperature and precipitation time series for US climate divisions. J Appl Meteorol Climatol 2014;53(5):1232–51.
[55] The American Society of Mechanical Engineers. Welded and Seamless Wrought Steel Pipe. ASME 2015. B36.10M-2015.
[56] Toprak S. Segmented pipeline damage predictions using liquefaction vulnerability parameters. Soil Dyn Earthquake Eng 2019;125:105758.
[57] Engelhardt MO. Rehabilitation strategies for water distribution networks: a literature review with a UK perspective. Urban Water 2000;2(2):153–70.
[58] Najafi M. Trenchless technology: pipeline and utility design, construction, and renewal. McGraw-Hill Education; 2005.
[59] Wols B, Van Daal K, Van Thienen P. Effects of climate change on drinking water distribution network integrity: predicting pipe failure resulting from differential soil settlement. Procedia Eng 2014;70:1726–34.
[60] Science US, Administration E. Soil taxonomy: a basic system of soil classification for making and interpreting soil surveys. US Department of Agriculture; 1975.
[61] Baracos A, Hurst WD, Legget RF. Effects of physical environment on cast-iron pipe. Journal (American Water Works Association) 1955;47(12):1195–206.
[62] Sattar AM, Gharabaghi B, McBean EA. Prediction of timing of watermain failure using gene expression models. Water Resour Manage 2016;30(5):1635–51.
[63] Harris, D.M., Digital Design and Computer Architecture. 2019.
[64] Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter 2004;6(1):20–9.
[65] Lundberg, S. and S.-.I. Lee, A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874, 2017.
[66] Benesty J. Pearson correlation coefficient, in noise reduction in speech processing. Springer; 2009. p. 1–4.
[67] Cramér H. Mathematical methods of statistics, 43. Princeton university press; 1999. Vol.
[68] Fisher RA. Statistical methods for research workers, in breakthroughs in statistics. Springer; 1992. p. 66–70.
[69] Lerman P. Fitting segmented regression models by grid search. J R Statis Soc 1980; 29(1):77–84.
[70] Strumbelj E, Kononenko I. An efficient explanation of individual classifications using game theory. J Machine Learn Res 2010;11:1–18.
[71] Mangalathu S, Hwang S-H, Jeon J-S. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. Eng Struct 2020;219:110927.
[72] Feng D-C. Interpretable XGBoost-SHAP Machine-Learning Model for Shear Strength Prediction of Squat RC Walls. J Struct Eng 2021;147(11):04021173.
[73] Wen X. Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP. Accident Anal Prevent 2021;159:106261.
[74] Singh A, Adachi S. Bathtub curves and pipe prioritization based on failure rate. Built Environ Project Asset Manag 2013.