Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

Full Length Article

Explainable ensemble models for predicting wall thickness loss of water pipes

Ridwan Taiwo [a], Abdul-Mugis Yussif [a], Mohamed El Amine Ben Seghier [b,*], Tarek Zayed [a]

[a] *Department of Building and Real Estate, The Hong Kong Polytechnic University, Hung Hum, Hong Kong*
[b] *Department of Built Environment, Oslo Metropolitan (OsloMet) University, Oslo, Norway*

ARTICLE INFO

Keywords:
Water pipelines
Wall thickness
Machine learning
Ensemble learning
SHAP

ABSTRACT

Water Distribution Networks (WDNs) are susceptible to pipe failures with significant consequences. Predicting wall-thickness loss in pipes is vital for proactive maintenance and asset management. This study develops optimized, explainable machine learning models for this purpose. Data from four WDNs located in Canada and the USA are collected and preprocessed. Decision Tree, Random Forest (RF), XGBoost, LightGBM, and CatBoost are employed, with optimized hyperparameters via Tree-Structured Parzen Estimator. The proposed framework performance is assessed using dissimilarity-based and similarity-based metrics. Hyperparameter optimization substantially enhances predictive performance such that the mean absolute error of RF improved by 20.51%. Based on the evaluation metrics, the Copeland algorithm was employed to rank the models, and CatBoost emerged as the best-performing model with a Copeland score of 4, followed by XGBoost and RF. The Taylor Diagram offers a visual representation of the linear proportionality between observed and predicted values across various models, with CatBoost and XGBoost showing strong alignment. SHAP analysis identifies age, diameter, and length as key contributors. The optimized models proactively identify potential pipe failures, enhancing maintenance and WDN management.

1. Introduction

Water Distribution Networks (WDNs) play a crucial role in ensuring reliable supplies of clean water to communities and industries. However, the deterioration and failure of these infrastructures can have significant consequences across the sustainability dimensions (i.e., economic, environmental, and social). Understanding the failure mechanism of water pipes is essential for developing effective predictive models to prevent such failures and ensure efficient WDNs management [1].

Regarding the economic consequences, pipe failures necessitate immediate repairs or replacement, which incur significant costs for materials, labor, and equipment [2]. For instance, a study conducted by Xu et al. [3] found that pipe failures in China resulted in a rehabilitation cost of more than 10 billion RMB in 2014. It is estimated that the United States needs to spend about $30 billion yearly to rehabilitate its WDNs [2]. Furthermore, pipe failures lead to water leakage, resulting in substantial water loss for utility providers. This loss not only affects the revenue stream but also increases operational expenses associated with water treatment and pumping. It is assessed that approximately 7 billion gallons of water are lost each year due to pipe failures in the United States alone, amounting to billions of dollars in economic losses annually [2,4].

As per the environmental consequences, water pipe failures increase energy consumption, as utilities need to pump and treat additional water to compensate for the losses [5]. This increased energy demand not only drives up operational costs but also leads to higher greenhouse gas emissions, contributing to climate change. Moreover, pipe failure leads to erosion, flooding, and traffic congestion [6]. Los Angeles experienced a major water pipe failure in 2014, where pipes within 80–90 years experienced a burst, causing extensive flooding and damage. The incident resulted in the loss of more than 20 million gallons of water, significant repair costs, disruption of traffic and businesses, and substantial environmental impact due to water wastage [7].

The social impacts of pipe failure include an increased risk of getting affected by waterborne diseases, public service disruption, and damages [8]. Failure of water pipes affects households, businesses, and critical infrastructures such as hospitals and schools [6]. Lack of access to clean water jeopardizes public health and sanitation, hindering daily activities and posing a risk during emergencies. In 2016, the city of Flint,

* Corresponding author.
E-mail addresses: ridwan-a.taiwo@connect.polyu.hk (R. Taiwo), mugis.yussif@connect.polyu.hk (A.-M. Yussif), moseg7662@oslomet.no (M.E.A. Ben Seghier), tarek.zayed@polyu.edu.hk (T. Zayed).

https://doi.org/10.1016/j.asej.2024.102630
Received 13 August 2023; Received in revised form 23 December 2023; Accepted 29 December 2023
Available online 16 January 2024

Nomenclature

AC	Asbestos Cement
ANN	Artificial Neural Network
ANFIS	Adaptive Neuro-Fuzzy Inference System
AUC	Area Under the Curve
BT	Boosted Trees
CE	Cementitious
CI	Cast Iron
Cox-PHM	Cox Proportional Hazard Model
C-Index	Concordance Index
CP	Cathodic Protection
CML	Cement Mortar Lining
DI	Ductile Iron
DT	Decision Tree
EL	Ensemble Learning
ELM	Extreme Learning Machine
EFB	Exclusive Feature Bundling
FIS	Fuzzy Inference System
GBDT	Gradient Boosting Decision Tree
GEP	Gene Expression Programming
GOSS	Gradient-based One-Side Sampling
GRNN	General Regression Neural Network
KGE	Kling-Gupta Efficiency
KNN	K-Nearest Neighbors
LightGBM	Light Gradient Boosting Machine
LR	Logistic Regression
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MARS	Multivariate Adaptive Regression Spline
ML	Machine Learning
MLP	Multi-Layer Perceptron
MLR	Multiple Linear Regression
MSE	Mean Squared Error
MTF	Mean Time to Failure
M5T	Model Tree
NNR	Non-Linear Regression
NSE	Nash-Sutcliffe Efficiency
OB	Ordered Boosting
OTS	Ordered Target Statistics
PE	Polyethylene
PVC	Polyvinyl Chloride
RAE	Relative Absolute Error
RF	Random Forest
RMSE	Root Mean Squared Error
RSF	Random Survival Forest
RUL	Remaining Useful Life
SD	Standard Deviation
SHAP	SHapley Additive exPlanations
SI	Scatter Index
SVC	Support Vector Classifier
SVM	Support Vector Machine
TD	Taylor Diagram
TPE	Tree-Structured Parzen Estimator
U95	95 % Uncertainty Interval
WDN	Water Distribution Network
WPHM	Weibull Proportional Hazard Model
WPHSM	Weibull Proportional Hazard Survival Model
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting
SMBO	Sequential model-based optimization
σ	Standard Deviation
μ	Mean
W_i	*i-th* measured value of wall thickness loss
P_i	*i-th* predicted value of wall thickness loss
$\bar{W}$	Arithmetic mean of the measured values of wall thickness loss
r	Correlation between the measure and predicted value of wall thickness loss
σ_{Pi}	Predicted values standard deviation
σ_{Wi}	Measured values standard deviation
μ_{P_i}	Mean of the predicted values
μ_{W_i}	Mean of the measured values
$\varnothing_i(f)$	Shapley value of feature i
S	subset of features

Michigan, USA, faced a severe water crisis when lead-contaminated water flowed through aging pipes, exposing residents to health risks. The incident highlighted the social consequences of water pipe failures, including public health emergencies and long-term health and developmental issues, particularly affecting vulnerable populations [9].

The statistics mentioned above provide compelling evidence of the significant issues caused by water pipe failures, highlighting the urgent need for attention and mitigation strategies. In the literature, various forms of failure indicators have been established, including the failure probability, failure rates, time-to-failure, remaining useful life, condition index, and wall-thickness loss [10–12]. While predictive models have been developed to forecast some of these failure indicators for water pipes, such as the failure probability, failure rates, time-to-failure, remaining useful life, and condition index, there is a noticeable gap in the literature regarding the prediction of wall-thickness loss specifically. Wall-thickness loss is directly related to water pipe failure. As pipes age and undergo deterioration, the gradual thinning of the pipe-wall due to corrosion, erosion, or other forms of material degradation can significantly impact the pipe's structural integrity [13,14]. The loss of wall-thickness reduces the pipe's ability to withstand internal and external pressures, increasing the risk of failure [15]. Wall-thickness loss is critical for assessing the remaining lifespan of water pipes, even though it has received relatively less attention in research efforts. The existing gaps in the literature are summarized as follows:

- The current literature lacks sufficient studies specifically addressing the prediction of wall-thickness loss in water pipes.
- Various existing prediction models lack a systematic approach for selecting and optimizing the hyperparameters of machine learning (ML) models.
- The majority of ML-models employed in pipe failure prediction are often considered black-box models, offering limited insights into their decision-making process.

Therefore, the aim of this study is to contribute to the existing knowledge about WDN by developing optimized ML models to predict the wall-thickness loss of water pipes. The specific objectives are highlighted below:

- To develop optimized ensemble models for predicting wall-thickness loss in water pipes.
- To compare and rank the optimized models using the Copeland algorithm.
- To interpret the best-optimized model using the SHapley Additive exPlanations (SHAP) technique, which will provide insights into the influential factors and the decision-making process of the model, enabling a better understanding of the failure mechanism of water pipes.

This study employs ensemble learning (EL), a powerful approach in ML, which has shown promising results in various domains [16,17] by combining multiple individual models to improve prediction performance increase robustness, and provide enhanced generalization capabilities. By harnessing the diversity of individual models, EL models can mitigate bias, reduce overfitting, and achieve superior predictive accuracy compared to single models. In addition to accurate predictions, there is an increasing demand for explainable artificial intelligence (XAI) models in critical infrastructure management. Hence, this study explains the contribution of the input variables to the predictive model by leveraging SHapley Additive exPlanations (SHAP). The outcomes of this research have significant practical implications for water utility managers and decision-makers, enabling them to make informed decisions regarding maintenance prioritization, rehabilitation planning, and resource allocation. Ultimately, this study contributes to the advancement of predictive modeling techniques for water infrastructure management, facilitating the reliable provision of clean water to communities and ensuring the long-term sustainability of WDNs.

2. Literature review

As indicated in the previous section, indicators used for predicting failure in water pipes include probability of failure, failure rates, time-to-failure, remaining useful life, condition index, and wall thickness loss. Table 1 summarizes the existing studies in this regard, including the used techniques, predicted failure indicator, adopted evaluation metrics, pipe type, data splitting ratio, and study location. In cases where more than one model is developed in a study, the evaluation metrics for the best model are reported.

Amiri-Ardakani & Najafzadeh [18] applied three algorithms to model the failure rate of water pipes in Yazd's WDN, located in Iran. These algorithms include multivariate adaptive regression spline (MARS), gene expression programming (GEP), and Model Tree (M5T). The pipes' diameter ranged from 63 mm to 110 mm. In terms of the adopted evaluation metrics (correlation coefficient (R) and RMSE), the MARS model outperformed the other two algorithms. For instance, the R for the MARS model was 0.981, while that of GEP and M5MT were 0.971

Table 1
Summary of the related studies for predicting the failure indicators for water pipes.

Reference	Technique	Failure indicator	Evaluation metrics	Type of pipes	Data splitting	Study location
[30]	LR	Probability of failure	AUC – 0.680 Recall – 0.672 Acc – 0.800	AC, CI, DI, PVC, and others	75 % Testing – 25 %	Austin, USA
[29]	XGBoost, RF, BT	Probability of failure	AUC = 0.8992	CI, DI, PVC, and others	Training – 12 years data Testing –3 years data	USA
[32]	ANN, LightGBM, LR, KNN, and SVC	Probability of failure	AUC – 0.81 Recall – 0.861	CI, DI, and others	Training – 80 % Testing – 20 %	Cleveland, USA
[27]	WPHSM, RF, and RSF	Remaining useful life	C-Index = 0.925	AC, CI, and DI	Training – 80 % Testing – 20 %	Canada
[10]	ANFIS and FIS	Condition index	R2 = 0.9145 RMSE = 0.6829	–	Training – 60 % Testing – 40 %	Arequipa, Peru
[18]	MARS, GEP, and M5 Tree	Failure rate	R = 0.981 RMSE = 0.544	AC, CI, PE	Training – 80 % Testing – 20 %	Yazd, Iran
[11]	LR and SVR	Probability of failure	AUC – 0.873 Recall – 0.848 Acc- 0.769	CE, PL, and ME.	Training – 5 years data Testing – 2 years data	Seville, Spain
[33]	ANFIS and ANN	Remaining useful life	MAE = 0.880 MAPE = 5.431 RAE = 0.007	AC, CI, DI, and Steel	Training – 75 % Testing – 25 %	USA and Canada
[24]	ANN, RF, and XGBoost	Time to failure	R – 0.85 RMSE – 5.81	AC, CI, DI, and PVC	Training – 80 % Testing – 20 %	North America
[21]	Extreme Learning Machine	Failure rate	R2 – 0.65 RMSE – 0.09	AC, CI, and DI	Training – 75 % Testing – 25 %	Toronto, Canada
[25]	ANN	Remaining useful life	R2 – 0.9877 MAE – 3.890 MAPE – 2.870	AC, CI, Concrete, DI, PE, PVC, Steel, and Copper	Training – 70 % Testing – 30 %	Quebec, Canada
[19]	ANN	Failure rate	R2 = 0.4142	CI, PE, PVC, and Steel	Training – 50 % Testing – 50 %	Poland
[20]	WPHM, Cox-PHM, and Poisson Model	Failure rate	RRSE – 0.31 MAE – 7.3 RMSE – 9.7	CI, DI, and PVC	Training – 70 % Testing – 30 %	Calgary, Canada
[23]	ANN	Time to failure	R – 0.82 RE – 0.32	AC, CI, DI, and PVC	Training – 70 % Testing – 30 %	Scarborough, Canada
[22]	Survival analysis	Time to failure	–	–	–	–
[26]	MLP, GRNN, and M.R.	Remaining useful life	R2 = 0.96 MAE = 0.12	CI	Training – 80 % Testing – 20 %	USA and Canada
[31]	ANN	Condition index	R2 = 0.8629	CI, DI, and Steel		South Korea

and 0.888, respectively. While the study contributed to the failure prediction rates of water pipes, the predictive capability of the model could be enhanced by systematically selecting the best hyperparameters for the models. Employing the data from 261 distribution pipes and 306 house connection pipes, Kutyłowska [19] established predictive models based on artificial neural network (ANN) for forecasting the failure rate of water pipes. Although the model achieved an R^2 of 0.9510 for the house connection pipes, the R^2 on the testing dataset was significantly low (i.e., $R^2 = 0.4142$). As the author suggested, the model's accuracy can be improved by incorporating more input variables. Using three statistical methods – The cox proportion hazard model (Cox-PHM), Poisson model, and Weibull proportional hazard model (WPHM) – Kimutai et al. [20] predicted the failure rate of pipes in a WDN in Calgary, Canada. The network is majorly dominated by polyvinyl chloride (PVC) pipes (54 %), followed by cast iron (CI) pipes (20.3 %) and ductile iron (DI) pipes (14.5 %). It was found that WPHM outperformed the other models. For instance, the relative absolute error (RAE) of the WPHM model using the CI pipe dataset was found to be 9.7, while that of the Cox-PHM and Poisson models was 16 and 11, respectively. Similarly, an extreme learning machine (ELM) has been employed to develop a failure rate predictive model [21]. Data was collected from a WDN in Toronto, Canada. The analysis data revealed that using pipe protection techniques such as cathodic protection (CP) and cement mortar lining (CML) reduced water pipes' failure rate by 60 % and 80 % for CI and DI pipes, respectively. In order to evaluate the predictive ELM performance, its efficiency was compared against alternative ML algorithms, including ANN, non-linear regression (NNR), and support vector machine (SVM). The results indicated that the ELM model exhibited superior performance compared to the other ML models, demonstrating its effectiveness in accurately predicting failure rates.

While establishing the model for condition assessment scoring of pipes, Opila & Attoh-Okine [22] calculated the mean time to failure (MTF) for pipes in a network. The MTF was estimated by integrating the failure probability over time. Subsequently, the condition grades of pipes were computed using the discounting process, which relies on the MTF value of the pipes. Using length, diameter, year of installation, number of previous breaks, and soil type as the input variables, Harvey et al. [23] developed ANN models to predict the time to failure of water pipes. The dataset employed in the study was obtained from a 5850 km WDN situated in Canada. While the network encompasses four material types (AC, CI, DI, and PVC), ANN models were exclusively developed for three of the materials due to the high imbalance observed data for the fourth material (PVC). The R and relative error (RE) of the DI model was 0.82 and 0.32, respectively. Similarly, a comparison between ANN, RF, and XGboost was made regarding the prediction of time-to-failure of water pipes [24]. The models were applied to the data of a WDN in a North American city. The results showed that XGboost outperformed the other two algorithms in predicting the time-to-failure.

The remaining useful life (RUL) of water pipes is another failure indicator that has been investigated in the extant literature. Zangenehmadar et al. [25] developed an ANN model based on the Levenberg-Marquardt backpropagation algorithm using five input variables (age, diameter, length, breakage rate, material, and condition) to forecast the RUL of water pipes. The ANN model achieved R^2, MAE, and MAPE values of 0.9877, 3.890, and 2.870, respectively. While the model accurately predicted the RUL of water pipes, the interpretability of the model remains a challenge. Furthermore, multi-layer perceptron (MLP), general regression neural network (GRNN), and multiple regression (MR) were employed by [26] to predict the RUL of cast iron pipes located in the USA and Canada. The data was collected from 16 municipalities in the two countries. 20 % of the data consisting of 136 pipes were used for model testing. As per the evaluation metrics, MLP outperformed GRNN and MP models with an R^2 and MAE of 0.96 and 0.12, respectively. In their study, Snider & McBean [27] aimed to estimate the RUL of water pipes by calculating the time to the next breakage. The authors defined the time to the next breakage as the point at which the pipe repairing cost exceeds the average replacement cost. In essence, this approach determined that pipe replacement is more cost-effective when the repair costs surpass the replacement cost. To estimate the time to the next breakage, the authors employed three different models: Random Survival Forest (RSF), Random Forest (RF), and the Weibull proportional hazard survival model (WPHSM). To evaluate the predictive performance of the models, the researchers utilized the concordance index (C-Index). This metric assesses the model's ability to rank the observed failure times correctly. Among the three models tested, the RSF model demonstrated the best predictive capability with the highest C-Index score.

In the study conducted by Robles-velasco et al. [28], the failure probability of pipes in a WDN located in Seville, Spain, was investigated. The WDN had a total length of 3800 km and consisted of pipes made of cementitious (CE), plastic (PL), and metallic (ME) materials. The dataset for model training ranged from 2012 to 2016, while the models were tested on data from 2017 to 2018. Logistic regression (LR) and support vector machine (SVM) were employed as predictive models to estimate the failure probability. The performance of these models was evaluated using metrics derived from the confusion matrix, such as recall and the area under the curve (AUC). It should be noted that the failure probability predictions need to be converted to binary values before establishing the confusion matrix. LR outperformed SVM based on the AUC metric, indicating better overall performance in predicting the failure probability. However, SVM also demonstrated promising results using the recall metric compared to LR. Chen et al. [29] utilized data from six utilities in the USA to develop models for predicting the failure probability of water pipes. Each of the utilities had different record durations; however, the overall common record period utilized in the study was from 2005 to 2018. The researchers employed three algorithms to develop the predictive models: XGBoost, RF, and Boosted Trees (BT). The findings of the study revealed that both XGBoost and RF models exhibited comparable performance, which was superior to the BT model. Similarly, Rifaai et al. [30] investigated the failure probability in a WDN using LR model, taking into account various factors, including pipe attributes, environmental conditions, operational factors, and failure history. To address the correlation between repeated observations for the same pipe, generalized estimating equations were utilized. The data were collected from a WDN located in Austin, USA. One of the challenges encountered in the analysis was the highly imbalanced nature of the dataset, with only 6.5 % of the pipes having experienced failure in the past. The evaluation metrics including accuracy, precision, recall, and area under the curve (AUC) were reported as 0.80, 0.67, 0.67, and 0.68, respectively.

Another failure indicator that has been investigated is the condition index of water pipes. Dawood et al. [10] presented a framework that combines the adaptive neuro-fuzzy inference system (ANFIS) and fuzzy inference system (FIS) to assess and predict the condition index of eight WDNs. The study utilized historical data from eight provinces in the Arequipa region of Peru to train and test the ANFIS model. This model was employed to calculate the network condition index for each province, capturing the unique characteristics and conditions of WDNs. To aggregate the individual province indices into a holistic representation of the region's network condition, the FIS model was utilized. The resulting network condition index for the Arequipa region was 63.1, reflecting the overall WDNs health and performance. The proposed framework was validated by comparing it with the multiple linear regression model (MLP) and showed better performance and accuracy. Geem et al. [31] utilized ANN to predict the condition index of pipes in a WDN located in South Korea. Due to data limitations, only 21 out of the 61 available records were considered for the analysis. To train the ANN model, 11 records were utilized, while the remaining 10 records were reserved for validation purposes. The evaluation of the ANN model on the validation dataset yielded an R^2 value of 0.8629. This indicates a reasonably good fit of the model to the observed data. However, it is important to note that the small size of the dataset used in the study may have led to overfitting.

The literature review presented above reveals gaps in the existing studies pertaining to predictive models for water pipes in WDN. The identified gaps are elaborated as follows:

1. **Limited Focus**: The current literature lacks sufficient studies specifically addressing the prediction of wall-thickness loss in water pipes. This gap highlights the need for more research efforts directed toward forecasting this critical aspect of pipe deterioration.
2. **Insufficient Hyperparameter Optimization:** Many existing prediction models lack a systematic approach for selecting and optimizing the hyperparameters of ML-models. This deficiency requires developing methodologies that effectively optimize the hyperparameters to enhance the predictive performance of the ML-models.

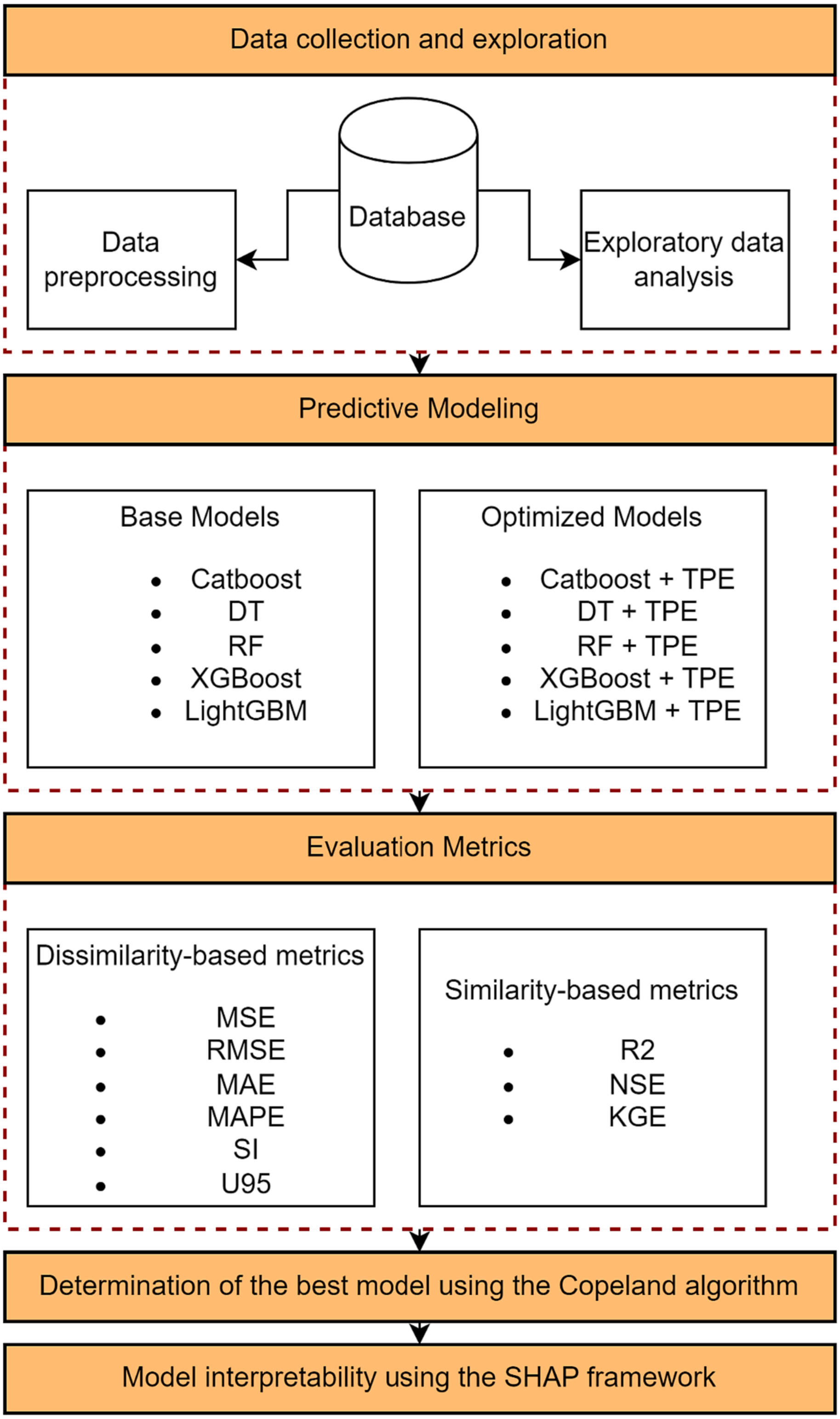

Fig. 1. The proposed framework for predicting wall-thickness loss of water pipes.

3. **Lack of Interpretability:** The majority of ML-models employed in pipe failure predictions are often considered black-box models, offering limited insights into their decision-making process. This gap highlights the necessity for research that focuses on the interpretability of ML-models to provide a deeper understanding of the factors influencing pipe failure.

3. Methodology and database

3.1. Methodology

The adopted framework for this study is illustrated in Fig. 1, consisting of five distinct phases. The first phase focuses on data collection and exploration. The data utilized in this study is obtained from previous experimental investigations conducted on four WDNs located in Canada and the USA, as documented by [34]. To ensure data quality, preprocessing techniques are applied, involving the removal of outliers and the conversion of input variables to the metric system (SI units). Explanatory data analysis is then conducted to gain insights into the characteristics and patterns within the dataset. In the second phase, predictive models are developed using various algorithms, including decision tree (DT), random forest (RF), extreme gradient boosting (XGBoost), light gradient boosting machines (LightGBM), and categorical boosting (CatBoost). To enhance the predictive capability of these base models, the Tree-Structured Parzen Estimator (TPE) is employed to optimize the hyperparameters. This approach aims to improve the models' performance and accuracy by fine-tuning their configuration.

In the third phase, the developed models are evaluated using both dissimilarity and similarity-based metrics. The dissimilarity-based metrics include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Scatter Index (SI), and 95 % Uncertainty Interval (U95). The similarity-based metrics encompass R-squared (R^2), Nash-Sutcliffe Efficiency (NSE), and Kling-Gupta Efficiency (KGE) [35,36]. By employing a comprehensive range of evaluation metrics, a holistic assessment of the models' performance is achieved. Subsequently, in the fourth phase, the Copeland algorithm is utilized to rank the models based on their evaluation metrics, enabling the selection of the best-performing model. This algorithm allows for a systematic comparison of the models, considering their relative performance across multiple metrics. The model with the highest rank is considered the most suitable for predicting water pipe wall thickness loss. Finally, in the fifth phase, the selected best model is interpreted using the SHapley Additive exPlanations. All computations in this study, including the development of the predictive models, are performed using Python 3.11 programming language and its associated libraries, such as Scikit-learn, Pandas, Seaborn, Matplotlib, XGBoost, LightGBM, CatBoost, SkillMetrics, and Shap [37].

3.2. Data collection and exploration

The data utilized in this study is obtained from WDNs in Canada and the USA, comprising six independent variables: age, length, diameter, material, number of breaks, and installation year. The dependent variable of interest is the wall-thickness loss, measured as a percentage. The database comprises 235 sample pipes, a size commonly found in experimental investigation databases [17]. To facilitate modeling, the categorical variable of material is preprocessed as dummy variables. The descriptive statistics of the numerical data are presented in Table 2, providing insights into each variable's central tendency and variability.

The descriptive statistics provide an initial understanding of the characteristics and range of the variables in the database. It is observed that the average age of the water pipes in the dataset is approximately 50.76 years, with a standard deviation of 32.21 years. The minimum and maximum ages recorded are 1.0 and 131.0 years, respectively. Moreover, Fig. 2 presents the correlation matrix of the dataset's variables. The correlation is computed using the Pearson Correlation Coefficient. The results reveal a moderate correlation between the majority of the independent variables and the dependent variable, wall thickness loss, as the correlation coefficients predominantly fall within the range of 0.3 to 0.5 [38].

Table 2
Descriptive statistics of the database.

Factor	Mean	Standard deviation	Minimum	Maximum
Age (Years)	50.76	32.21	1.0	131.0
Length (m)	986.29	1604.71	6.25	11029.27
Diameter (m)	0.28	0.13	0.10	0.60
Number of breaks	5.32	8.05	1.00	95.00
Installation year	1959.45	30.31	1887.00	2011.00
Wall thickness loss (%)	23.57	13.02	1.00	49.00

Fig. 3 presents the histograms of the numerical data. The independent variables are shown in orange, while the dependent variable is presented in green for easy visualization. The visualization of the distributions aids in identifying any potential outliers, as seen in the histogram of the “number of breaks” variable, which are data points that significantly deviate from the rest and could impact the modeling process. To ensure data quality, preprocessing techniques were applied, including the removal of outliers using the interquartile range (IQR) method. The IQR represents the difference between the 75th and 25th percentiles of the data. Any data points that were below the 25th percentile minus 1.5 times the IQR, or above the 75th percentile plus 1.5 times the IQR, were removed from the dataset as potential outliers. This process eliminated erroneous and anomalous measurements that could adversely impact the modeling. Regarding the material distribution, 52.8 % is made up of CI, while 20.9, 16.2, and 10.2 % are made up of DI, CI, and steel, respectively.

4. Model development

This section explains the development of ensemble learning models for predicting the wall-thickness loss. It details the algorithms that are employed, including predictive, optimization, ranking, and interpretability algorithms.

4.1. Predictive modeling using decision tree

Decision trees (DTs) operate heuristically and can predict continuous values in regression tasks [39]. It is a tree-like algorithm comprising a hierarchy of nodes and leaves representing decisions and their possible consequences. The tree is constructed by splitting a root node into a few initial internal nodes, which are further divided into subsequent internal nodes based on the information-gain of an attribute feature (see Figs. 4 and 6). Nodes that terminate the decision trees are called terminal nodes and possess the maximum homogeneity of the decision class [40]. When using a DT for multivariate regression tasks, the goal is to minimize the difference between the parent node variance and the weighted sum of the child node variance.

A differential entropy $H(\alpha)$ is adopted to measure the random uncertainty that follows a continuous probability distribution in a multivariate regression (see Equation 1) (Cover & Thomas, 2005). That is, given $y = \mathbb{R}^d$, the learning goal is to generate a prediction model $\mathscr{M}(y|x)$ through a decision tree evaluation and storing a simple density model $\rho_l(y)$ at the leaf l.

$$\mathrm{H}(\alpha) = -\int_{\mathrm{y}} \alpha(\mathbf{y}|\mathbf{x})\log\alpha(\mathbf{y}|\mathbf{x})\mathrm{dy} \tag{1}$$

where $\boldsymbol{\alpha}(y|x)$ is a density with an absolutely continuous cumulative distribution function, and $\boldsymbol{\alpha}(x, y)$ is the true generating distribution. Optimizing the differential entropy of unknown continuous distribution

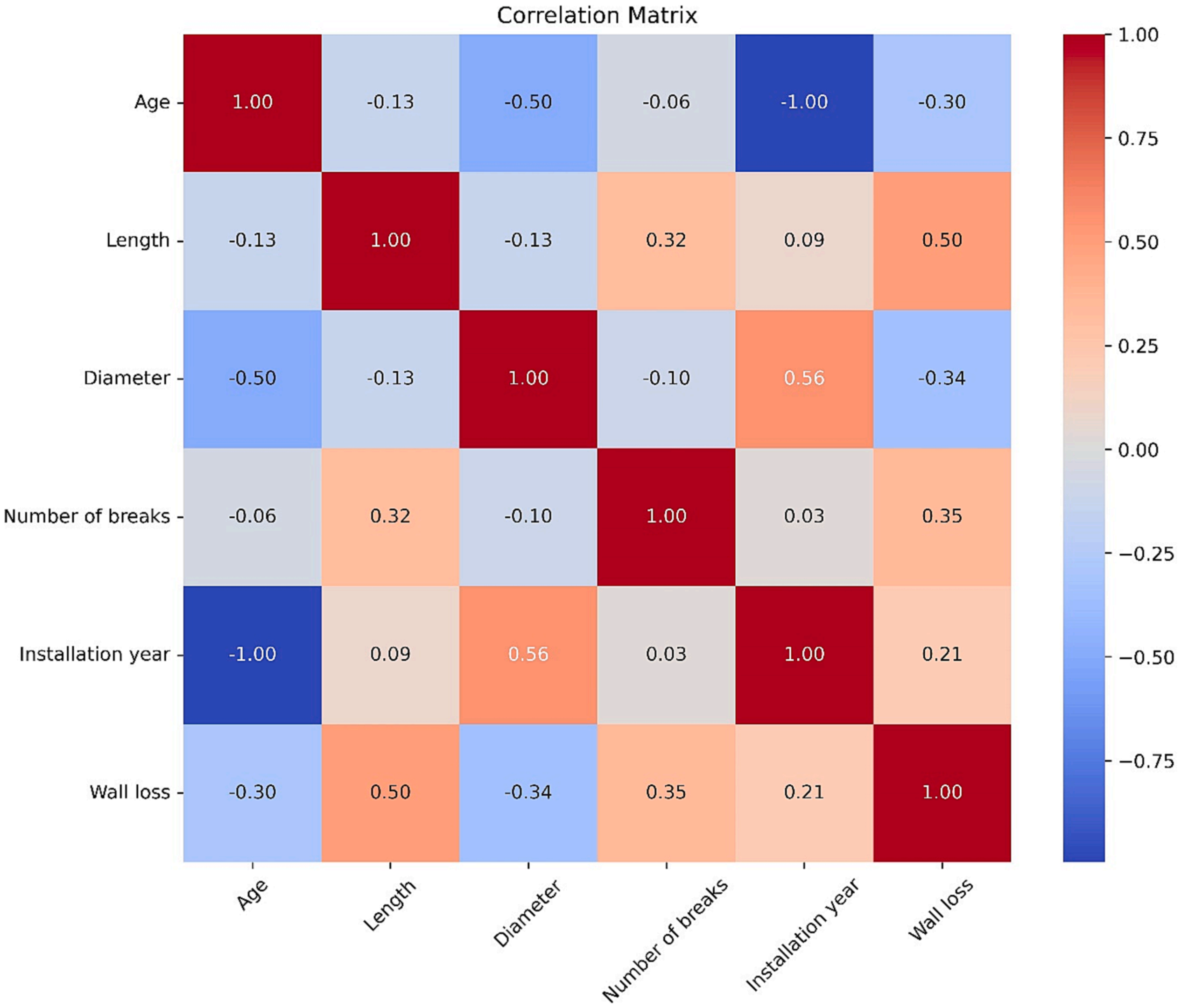

Fig. 2. Correlation matrix of the variables in the dataset.

is complicated; therefore, several estimators, including kernel density estimates, minimum spanning tree length, and *k*-nearest neighbor distances [41] have been proposed to achieve the maximum leaf homogeneity. DTs are the foundation for ensemble tree-type learning algorithms such as RF, CatBoost, XGBoost, and LightGBM.

4.2. Predictive modeling using Random Forest

Random Forest (RF) works on the concept of bootstrapping, i.e., by selecting random samples $S_1(x,y),\cdots,S_k(x,y)$ from the defined dataset with replacement [42]. It is inspired by bagging and feature randomness of the independent variables. A DT is built from each sample, considering only a random subset of the features at the node-splitting stage. This selection is contrary to what happens in the regular DT method, where all features are considered during partitioning. The final output of regression trees is decided by the mean of all the individual tree predictions [37]. Each tree is constructed to minimize the MSE of the prediction, as shown in Equation 2.

$$\mathrm{MSE(S)} = \frac{1}{|\mathrm{S}|}\sum\nolimits_{\mathrm{i}\in \mathrm{S}} (\mathbf{y_i} - \bar{\mathbf{y}}_\mathbf{S})^2 \tag{2}$$

where S is the sample space of the node, y_i is the individual target values, and $\bar{y}_S$ is the mean target value of the node. The MSE also serves as the splitting criterion for the regression model, whereas the Gini index determines how a node should be divided for the classification tasks [42]. Similar to DT, splitting starts with the root node (see Fig. 4), and the parent node splits to generate left and right branches of child nodes based on the former's mean squared residual.

4.3. Predictive modeling using Boosting Algorithms

4.3.1. XGBoost

Gradient Boosting Decision Trees (GBDT) are models that operate by fitting additive models in a forward stage-wise manner. A gradient-boosting model relies on the continuous learning of base learners to improve its performance, i.e., a weak learner is added to the present model at every stage to reinforce its learning capacity by reducing the prediction losses [43]. XGBoost learns additively from weak learners to strengthen its learning capacity and robustness. Due to its inherent parallel handling ability, it is scalable and efficient in handling large-scale classification and regression tasks. When dealing with real-world data, it does not require significant preprocessing due to the efficiency in managing missing data issues. XGBoost intends to minimize the objective function indicated in Equation 5. Since the gradients for all instances offer significant information to the GBDT algorithms, an instance that produces a small gradient possesses low deviation and is therefore considered well-trained. If there are K base learners $(\mathrm{f_k})$ to create an XGBoost model $(\bar{y}_i)$, f_k is successively added to $\bar{y}_i$ each time, as

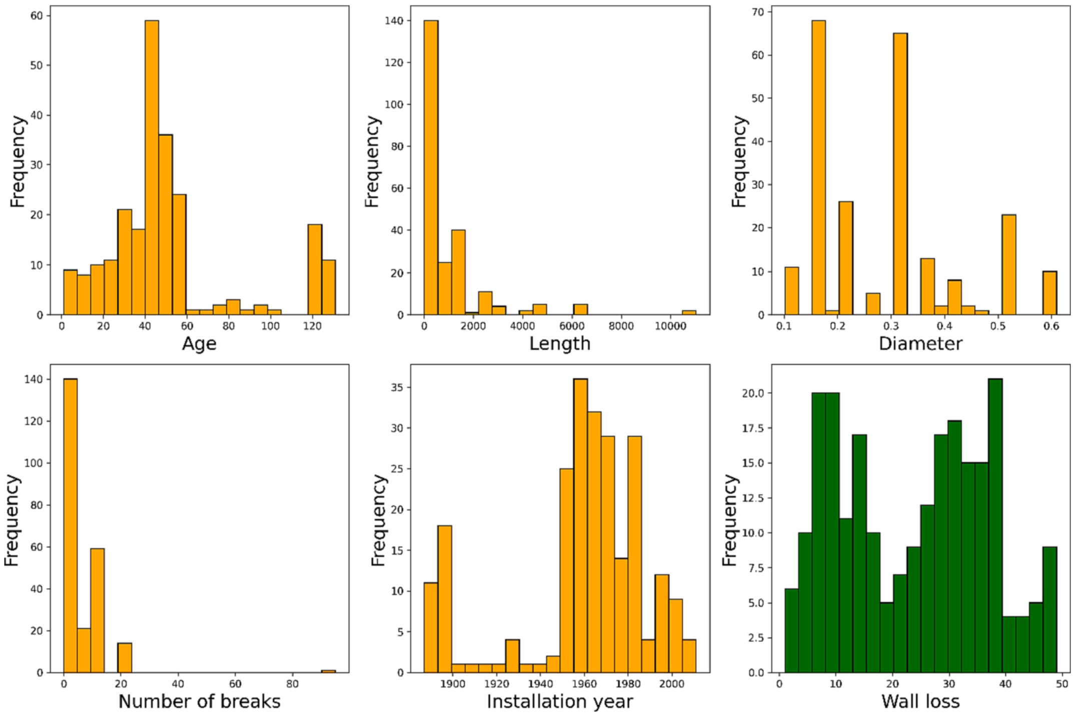

Fig. 3. Histograms of the numerical variables in the dataset.

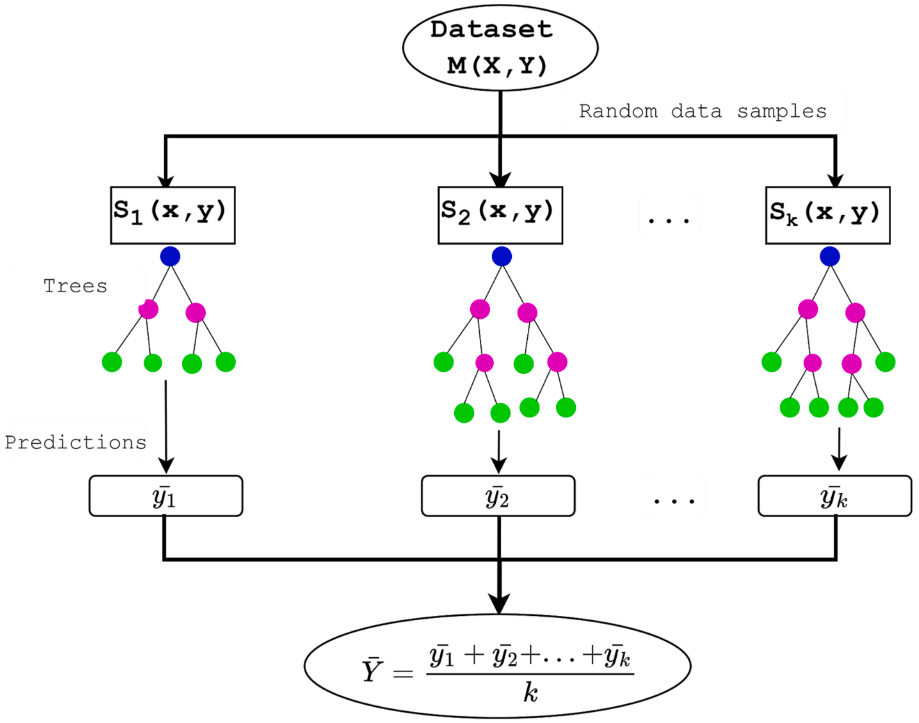

Fig. 4. The flowchart of the Random Forest approach.

shown in Equation 3 [44].

$$\bar{y}_i = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i), f_k \in F, \tag{3}$$

The function F is a space containing all the regression trees in the XGBoost model. For the full understanding of the functions that make up the model, the regularized objective function is minimized as shown below:

$$\mathscr{L} = \sum\nolimits_i \ell(\bar{y}_i, y_i) + \sum\nolimits_k \Phi(f_k)\Phi(f) = \gamma N + \frac{1}{2}\lambda||\omega||^2 \tag{4}$$

Optimizing the ensemble function is complex because it is trained additively; therefore, a greedy mechanism is used to achieve the desired optimization. Thus, f_t is the greedy function added to Equation 4 to minimize the model's loss at every iteration, such as the loss at Equation

5 [44].

$$\mathscr{L}^{\mathrm{t}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \ell[(\mathbf{y}_{\mathbf{i}}, \bar{\mathbf{y}}_{\mathbf{S}})^{(\mathrm{t}-1)} + \mathrm{f}_{\mathrm{t}}(\mathbf{x}_{\mathbf{i}})] + \Phi(\mathrm{f}_{\mathrm{t}}) \tag{5}$$

where $\bar{y}_S^{(t)}$ represents the i-th instance prediction at the t-th iteration. The regularized objective function has three parts: ℓ obtains the difference between the real values y_i and the predicted values $\bar{y}_S$, f_t is included to help minimize the function, and Φ manages the regression tree's complexity. N is the number of tree leaves, ω_i is the score or weight on the i-th leaf [44]. For regression tasks, as in this case, the scores of the leaves are continuous and shared by all the trees. XGBoost has shown high scalability and superior performance in many research areas, including infrastructure management and business domain [45]. Refer Fig. 5 for the graphical workflow of the XGBoost approach.

4.3.2. LightGBM

LightGBM is designed to reduce resource consumption while achieving high efficiency in speed and accuracy, especially for large-scale data tasks [46]. It employs two powerful techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to manipulate the gradient feature behavior for better scalability and efficiency. Normally, the instances with larger gradients contribute the most to the model gradient reduction. GOSS capitalizes on this feature and preserves those with larger gradients to whom it assigns larger weights while randomly sampling the instances with lower gradients. EFB uses the idea of feature exclusiveness to reduce the feature size for improved speed and memory consumption through feature bundling. Feature exclusiveness implies that it is often uncommon for most features to take nonzero values simultaneously since feature-exclusive values have very low occurrence rates [47].

LightGBM grows its trees leaf-wise (see Fig. 6) by preferring the most promising leaves, i.e., those with the maximum delta loss. The maximum delta loss is the largest loss reduction achieved when a node is partitioned [48]. For growth at the same leaf, leaf-wise growth algorithms have greater loss reduction than level-wise algorithms. Experimental results proved LightGBM (using GOSS) to exceed GBDT with the Stochastic Gradient Boosting in performance [47]. The quick implementation makes it ideal for practical applications of real-life problems because it can be updated frequently. The model's complexity at each computational stage is determined to decrease with the EFB adoption. If A is the given dataset, a sample data B is generated through bundling, such that the computational cost of building a histogram is reduced from ∇A to ∇B where $B \lll A$ [46].

4.3.3. CatBoost

CatBoost is a scalable ensemble learning technique developed for tackling categorical and numerical tasks with heterogeneous data and complex dependencies [49]. It adopts two innovative mechanisms: Ordered Target Statistics (OTS) and Ordered Boosting (OB). The OTS is a target-based encoding introduced to handle high cardinality features, which are not managed by the one-hot encoding in CatBoost implementation. Assuming a regression task to be performed on the dataset $\mathscr{M}(x_1, y_2), \cdots, (x_k, y_k)$ using the novel CatBoost algorithm, if h^t is the initial decision tree formulated. Then h^{t+1} is the decision tree to be added to form an ensemble in a way to minimize the loss $\mathscr{L}$. Therefore, the goal is to obtain a sample S from $\mathscr{M}$, that will bring about minimum losses [50].

However, it is crucial to ensure that overfitting would not occur from the encoding process using future data unavailable at the prediction time, a phenomenon known as target leakage. To avoid this, the OTS uses random permutations of the dataset to generate the required sample S for creating the required h^{t+1}. Then, the entire dataset $\mathscr{M}$ is used to evaluate whether the loss at $\mathscr{L}(y, F^t + h^{t+1})$ is minimized [50,51]. Additionally, it applies smoothing techniques to reduce the noise and variance of mean values in the training data, especially for low-frequency categories, to avoid imbalance or low-quality situations.

OB, on the other hand, involves how the ensemble is constructed by the addition of a new tree to CatBoost model without causing prediction shifts. It ensures that the model uses all the datasets (x_k, y_k) by the end of the training process for its generalizability requirements. Employing the OB is necessary as the initial application of random data selection by the previous techniques, and even the choice of data samples for building new trees h^{t+1}. Here, a new independent dataset S_k is always sampled from the entire dataset at each step of the boosting. It is then used as the new training examples for the current model, which was trained from the previously samples dataset S_{k-1} to obtain unshifted residuals [49]. Refer to Fig. 7 for the flowchart of the CatBoost approach.

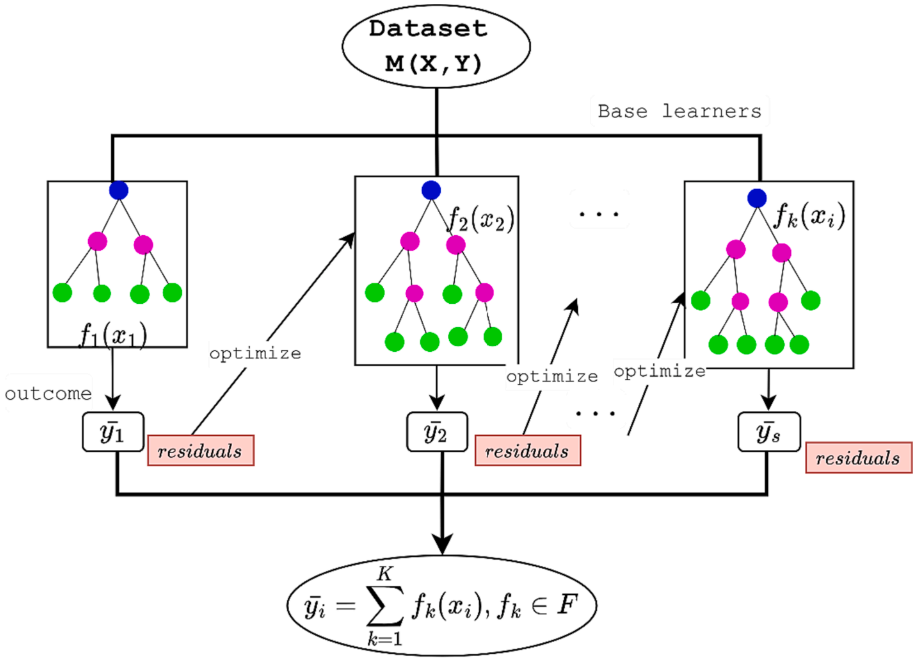

Fig. 5. The workflow of XGBoost approach.

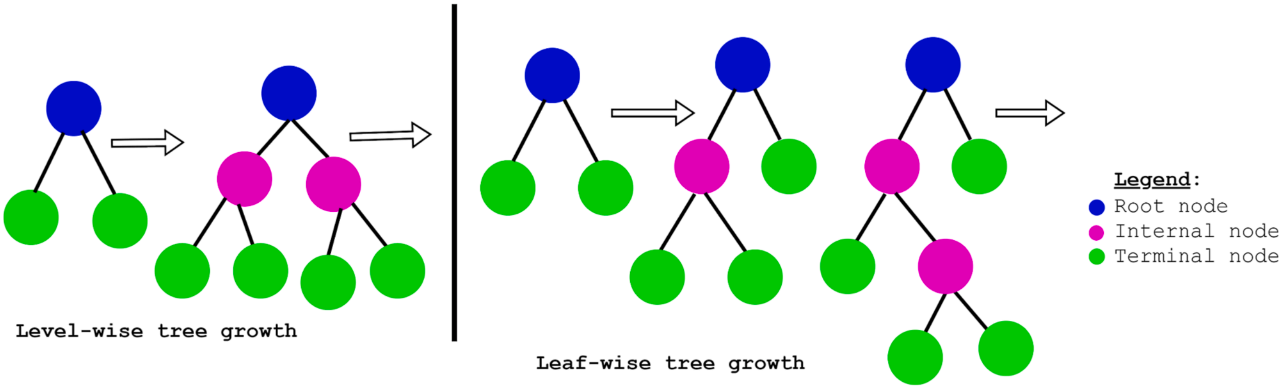

Fig. 6. Level-wise tree growth (Left) and Leaf-wise tree growth (Right).

4.4. Hyperparameters optimization using TPE

The Tree-structured Parzen Estimator (TPE) is a sequential model-based optimization (SMBO) approach that is widely used for hyper-parameter optimization in ML-models [52,53]. The TPE algorithm is a Bayesian optimization method that models $P(x|y)$ and $P(y)$ instead of $P(y|x)$ as in other SMBO methods. Here, x represents the hyper-parameters and y denotes the associated loss. The TPE algorithm operates by constructing a probabilistic model that maps hyperparameters to a probability of a score on the objective function. The algorithm iteratively refines this model as it gathers more data, using the model to select the most promising hyperparameters to evaluate on the actual

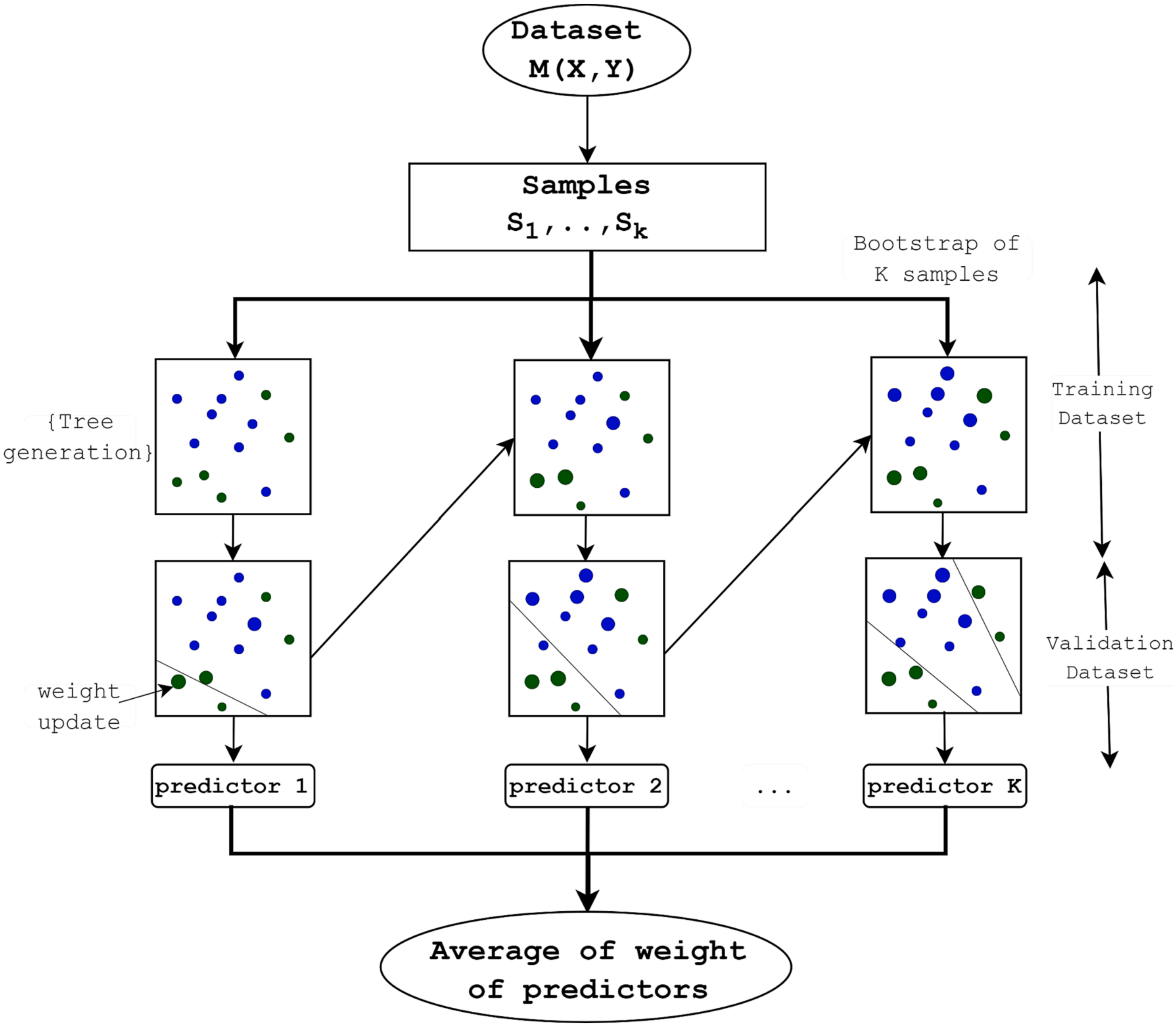

Fig. 7. The flowchart of CatBoost approach.

objective function [54].

Mathematically, the TPE algorithm defines two distributions: $l(x)$ for the best observed hyperparameters and $g(x)$ for the rest. The algorithm selects the next set of hyperparameters to evaluate the objective function by maximizing the Expected Improvement (EI) criterion, which is given by Equation 6 [55].

$$\mathrm{EI.}\,(\mathbf{x}) = \mathrm{E}\left[\max(\mathrm{f}(\mathbf{x}) - \mathrm{f}(\mathbf{x}^{*}),\, 0\,)\,\right] \tag{6}$$

where x is the point being evaluated, $\boldsymbol{x}^{*}$ is the current best point, $\mathrm{f}(x)$ is the predicted value of the function at $\boldsymbol{x}$, and E[.] denotes expectation.

The TPE algorithm follows these steps:

- Initialize by randomly sampling a single hyperparameter configuration and evaluating the MSE as the objective function value.
- Construct non-parametric Parzen estimator density models for the distributions $\mathrm{l}(\mathbf{x})$ and $\mathrm{g}(\mathbf{x})$. $\mathrm{l}(\mathbf{x})$ models the density of hyperparameter configurations that led to poorer objective scores. $\mathrm{g}(\mathbf{x})$ models the density of configurations with better objective scores.
- Generate a new candidate set of hyperparameters by sampling from $\mathrm{g}(\mathbf{x})$ and low-density regions of $\mathrm{l}(\mathbf{x})$. The expected improvement acquisition function is used to balance exploration and exploitation.
- Evaluate the MSE for the new hyperparameters and add them to the dataset.
- Rebuild the Parzen estimator models $\mathrm{l}(\mathbf{x})$ and $\mathrm{g}(\mathbf{x})$ using the updated data.
- Repeat steps 3–5 for M iterations until convergence criteria are met. Convergence is determined when no improvement in best MSE is observed over N iterations or maximum iterations M is reached.
- Return the hyperparameter configuration with the lowest MSE as the final solution.

The TPE algorithm has been shown to outperform other SMBO methods on a variety of benchmark problems, and thus, it was selected in this study [52,55]. Table 3 shows the hyperparameters of the five algorithms that were optimized, including their type and range.

4.5. Model evaluation metrics

In the assessment of the predictive EL-models, nine statistical metrics are employed to quantify their performance. These metrics are categorized into two groups: dissimilarity and similarity metrics [16]. The dissimilarity metrics include MSE, RMSE, MAE, MAPE, SI, and U95. The similarity metrics are R^2, NSE, and KGE. Thus, the dissimilarity metrics measure the disparity between predicted and actual values, while the similarity metrics assess the degree of resemblance between predicted and observed values [16]. The R^2 measures the proportion of variance in the dependent variable explained by the predictors, with values ranging from 0 to 1 and higher values denoting better model fit. NSE assesses how well the model replicates observed data by comparing residual and measured data variance; and ranges from $-\infty$ to 1. KGE provides a more comprehensive assessment by decomposing model performance into correlation, variability bias, and mean bias. This metric also ranges from $-\infty$ to 1, with higher KGE values suggesting stronger agreement between modeled and observed data [16,17]. Table 4 depicts the evaluation metrics and their corresponding formulas [36,56,57].

Table 3
Hyperparameter details of the EL-models.

Ensemble algorithms	Hyperparameters	Type	Range
DT	Maximum depth	Integer	[1,20]
	Minimum samples leaf	Integer	[2,20]
	Minimum samples split	Integer	[1,20]
RF	Maximum depth	Integer	[1,20]
	Minimum samples leaf	Integer	[2,20]
	Minimum samples split	Integer	[1,20]
	Number of estimators	Integer	[1, 1000]
XGBoost	Learning rate	Continuous	[0.01, 1]
	Maximum depth	Integer	[1,20]
	Number of estimators	Integer	[1, 1000]
	Subsample	Continuous	[0.01, 1]
LightGBM	Learning rate	Continuous	[0.01, 1]
	Maximum depth	Integer	[1,20]
	Number of estimators	Integer	[1, 1000]
	Subsample	Continuous	[0.01, 1]
CatBoost	Learning rate	Continuous	[0.01, 1]
	Maximum depth	Integer	[1,20]
	Number of estimators	Integer	[1, 1000]

4.6. Model ranking using the Copeland algorithm

In this study, the Copeland algorithm is used to rank various predictive models based on their performance. The algorithm works by comparing each model to every other model one by one. Points are given to the model that performs better in each comparison based on selected evaluation metrics [58].

(a) Performance Metrics: Criteria for Evaluation

In the scope of this research, multiple evaluation metrics serve as the fundamental criteria for ranking. For instance, metrics based on dissimilarity are evaluated using the principle:

$$LowerMSEValue \Rightarrow BetterModelFit \tag{7}$$

Conversely, similarity-based metrics are assessed by:

$$HigherR^{2}Value \Rightarrow GreaterExplanatoryPower \tag{8}$$

(b) Methodology of Pairwise Comparisons

The basis of the Copeland algorithm is the pairwise comparison of each model against every other model in the set. For a given pair (Model *i* and Model *j*), the evaluation metrics are compared, with points allocated based on the following criterion:

$$IfMetric_{i} > Metric_{j}, thenPoint_{i} + \; = 1 \tag{9}$$

$$IfMetric_{j} > Metric_{i}, thenPoinj + \; = 1 \tag{10}$$

Table 4
Evaluation metrics for the predictive models.

Performance indicator	Category	Formula	Remark
MSE	Dissimilarity-based	$\frac{1}{n}\sum_{i=1}^{n}(W_i - P_i)^2$	The lower the values of these metrics, the better the model
RMSE		$\sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(W_i - P_i)^2}$	
MAE		$\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\lvert W_i - P_i\rvert$	
MAPE		$\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\lvert\frac{W_i - P_i}{W_i}\rvert$	
SI.		$\frac{RMSE}{\bar{W}}$	
U95		$1.96\sqrt{(SD^2 - RMSE^2)}$	
R2	Similarity-based	$1 - \frac{\sum_{i=1}^{n}(W_i - P_i)^2}{\sum_{i=1}^{n}(W_i)^2}\; 0 \leq R2 \leq 1$	The closer the values of these metrics to 1, the better the model
NSE		$1 - \frac{\sum_{i=1}^{n}(W_i - P_i)^2}{\sum_{i=1}^{n}(W_i - \bar{W})^2} -1 \leq NSE \leq 1$	
KGE		$1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{Pi}}{\sigma_{Wi}} - 1\right)^2 + \left(\frac{\mu_{P_i}}{\mu_{W_i}} - 1\right)^2}$	

Note: W_i is the *i-th* measured value of wall-thickness loss, P_i is the *i-th* predicted value of wall-thickness loss, $\bar{W}$ is the arithmetic mean of the measured values of wall-thickness loss, r is the correlation between the measure and predicted value of wall thickness loss, σ_{Pi} is the predicted values standard deviation, σ_{Wi} is the measured values standard deviation, μ_{P_i} is the mean of the predicted values, and μ_{W_i} is the mean of the measured values.

This iterative process is conducted across nine evaluation metrics, concluding when all model pairs have undergone comparison.

(c) Aggregated Scoring and Quantification

Following pairwise assessments, an aggregate score, *S*, for each model is calculated using:

$$S = \sum Pointsgainedinallpairwisecomparisions \tag{11}$$

This aggregated score encapsulates the model's performance across the spectrum of comparisons, serving as a comprehensive metric of its relative effectiveness. Five different outcomes are recorded for each model: 'points,' 'wins,' 'losses,' Copeland score, and rank.

- **Points**: These are accumulated points given to a model when it performs better than another model in specific metrics during a pairwise comparison. For instance, if Model A has a lower MSE than Model B, Model A gains a score point. Scores are summed up across all comparisons to provide an aggregate score for each model.
- **Wins**: A 'win' is counted for a model when it accumulates more score points in a pairwise comparison than its competitor. For example, if Model A gains more points than Model B across all selected metrics in a particular pairwise comparison, then Model A is said to 'win' that comparison.
- **Losses**: Conversely, a 'loss' is recorded for a model when it gains fewer points than another model in a pairwise comparison.
- **Copeland score**: The Copeland score is calculated as the difference between the number of 'wins' and 'losses' for each model. Mathematically, it can be represented as:

$$CopelandScore = Wins - Losses \tag{12}$$

- **Rank**: Finally, the models are ranked based on their Copeland scores. The model with the highest Copeland score is ranked first, thereby indicating its superior performance across the evaluation metrics.

4.7. Model interpretability using SHAP framework

In this study, the SHAP framework is utilized to achieve model interpretability and gain insights into the wall-thickness loss prediction model. It should be noted that the best model selected by the Copeland algorithm is interpreted. The SHAP framework is based on cooperative game theory and provides a unified and mathematically grounded approach to explain the output of complex ML-models [59]. It assigns a unique value, known as the Shapley value, to each input feature, representing its contribution and direction to the prediction. The Shapley value ensures a fair and consistent allocation of contributions across features, considering all possible feature combinations.

Mathematically, the Shapley value for feature *i* is defined by Equation 7 [60]:

$$\varnothing_i(f) = \sum [p(S \cup \{i\}) - p(S)] \tag{7}$$

where $\varnothing_i(f)$ is the Shapley value of feature *i* for the prediction function *f, S* is a subset of features, excluding feature *I, p(S)* is the model's prediction when considering only the features in subset *S,* $p(S \cup \{i\})$ is the model's prediction when including feature *i* in the subset *S*. The Shapley value satisfies desirable properties, such as efficiency, linearity, and symmetry, making it an attractive method for feature attribution in ML-models.

5. Model implementation and discussion

This section discusses the results of the EL-models in relation to the base and optimized models, the selection of the best model, and its interpretability. It should be noted that 80 % of the data are used for training, while the remaining 20 % is employed for testing the models.

5.1. Results of the base EL-models

Table 5 reports the evaluation metrics of the base EL-models for predicting the wall-thickness loss of water pipes using the training and testing datasets. Variations in the model's performance across different metrics are observed, which highlights the importance of thorough evaluation. The five EL-models generally exhibit good performance on the training dataset, as indicated by the relatively low values for MSE, RMSE, MAE, and MAPE. Lower values for these metrics suggest that the models can closely predict the wall-thickness loss values. The high R^2, NSE, and KGE values further confirm the models' ability to explain the variance and achieve a strong fit to the training data. Using the testing datasets, which represent unseen data, the models show a decline in performance compared to the training datasets. This drop in performance is expected since the models encounter new patterns using data points not exposed to during training. Despite this drop, the EL-models still maintain reasonable predictive accuracy, as indicated by the relatively low values of the dissimilarity-based metrics and high values of the similarity-based values. When comparing the performance of the base EL-models, it is observed that each model has its strengths and weaknesses. The DT model shows exceptional performance on the training datasets, achieving the lowest MSE, RMSE, and MAE values among all EL-models. However, it shows a relatively higher drop in performance on the testing datasets, showing potential overfitting during the training phase. On the other hand, CatBoost, XGBoost, and RF consistently perform well across both datasets, achieving competitive results with relatively low MSE, RMSE, MAE, and MAPE values. CatBoost and XGBoost also exhibit high R^2 and NSE values, suggesting a strong fit on both datasets. RF also demonstrates consistent and competitive performance across both datasets.

5.2. Results of the optimized EL-models

The evaluation metrics for the optimized EL-models, as presented in Table 6, provide an assessment of their performance based on both training and testing datasets. The optimized EL-models demonstrate improvements in prediction accuracy during the training and testing phases compared to their respective base EL-models. For example, the RF model exhibited notable improvements, with MSE, RMSE, MAE, and MAPE decreasing by 19.56 %, 10.31 %, 17.70 %, and 6.93 %, respectively, during the training phase after hyperparameter optimization using TPE. A similar trend was observed during the testing phase, where the RF model's MSE, RMSE, MAE, and MAPE improved by 3.65 %, 1.84 %, 15.80 %, and 9.05 %, respectively, after optimizing the process. Furthermore, the optimized EL-models generally outperformed their base EL-models concerning similarity-based metrics such as R^2, NSE, and KGE. The R^2 values for the optimized EL-models were found to be very close to the unit, indicating a strong agreement between the measured and predicted values. For example, XGBoost model achieved an R^2 value of 0.9825 using the training data, while the optimized XGBoost model improved this performance to 0.9965. Similarly, using the testing data, the R^2 value increased by 5.30 %. The NSE and KGE metrics also reflected the enhanced performance of the optimized EL-models. NSE values approaching 1 indicate superior accuracy, which is demonstrated in NSE values compared to the base EL-models. Additionally, the KGE metric of XGBoost, which quantifies the agreement between EL-model predictions and observed values, improved by 3.96 % after optimization using TPE.

Moreover, in Fig. 8, scatter plots of the predicted versus measured values of the wall-thickness loss are presented using five optimized EL-models for both datasets. The red-line represents the best-fitting line for the measured values, while the green-line depicts the best-fitting line for the predicted values. Notably, when using the testing data, a remarkable agreement between the predicted and measured values of wall-thickness loss is observed, as evidenced by the minimal distance between the red and green lines, particularly for the optimized CatBoost and XGBoost

Table 5
Evaluation metrics for the base EL-models using the training and testing datasets.

Dataset	Models	MSE	RMSE	MAE	MAPE	SI	U95	R2	NSE	KGE
Training	DT	**0.5611**	**0.7491**	**0.2074**	**1.3133**	**3.0391**	**1.4682**	0.8530	**0.9966**	**0.9975**
	RF	5.8671	2.4222	1.9834	8.9891	8.8763	4.9721	0.9689	0.9700	0.9539
	XGBoost	0.5719	0.7562	0.2667	1.6550	3.0681	1.4822	0.9825	0.9903	0.9961
	LightGBM	14.8788	3.8573	2.9318	16.1763	15.6493	7.5603	0.9094	0.9099	0.9168
	CatBoost	2.4944	1.5793	1.2107	7.1171	6.4076	3.0955	**0.9848**	0.9849	0.9716
Testing	DT	28.2780	5.3177	3.9212	27.9766	27.5423	10.4210	0.8277	0.8314	0.9138
	RF	18.5690	4.3092	3.8305	26.4381	23.3914	9.9826	0.8459	0.8491	0.9199
	XGBoost	18.8455	4.3411	3.3708	**25.2200**	22.3700	8.4641	0.8552	0.8816	0.9085
	LightGBM	23.0591	4.8019	3.9197	37.2570	24.8738	9.4113	0.8513	0.8625	0.9182
	CatBoost	**16.2245**	**4.0279**	**3.1922**	29.8229	**20.8266**	**7.8800**	**0.9011**	**0.9032**	**0.9217**

*Bold numbers indicate the best result.

Table 6
Evaluation metrics for the optimized EL-models using the training and testing datasets.

Dataset	Models	MSE	RMSE	MAE	MAPE	SI	U95	R2	NSE	KGE
Training	DT + TPE	17.8790	4.2283	3.2676	16.1046	17.1547	8.2875	0.8912	0.8918	0.9208
	RF + TPE	4.7192	2.1723	1.6323	8.4234	8.8116	4.2569	0.9712	0.9714	0.9553
	XGBoost + TPE	0.5703	0.7518	**0.2181**	**1.3622**	**3.0474**	**1.4800**	**0.9965**	**0.9966**	**0.9973**
	LightGBM + TPE	18.4813	4.2989	3.2923	18.1405	17.4412	8.4260	0.9075	0.8881	0.8942
	CatBoost + TPE	**0.5424**	**0.7364**	1.1824	6.9775	6.2093	2.9997	0.9857	0.9858	0.9733
Testing	DT + TPE	19.3884	4.4032	3.1791	26.2625	22.8061	8.6289	0.8812	0.8844	0.9312
	RF + TPE	17.8909	4.2297	3.2250	24.0459	21.9107	8.2902	0.8910	0.8933	0.9302
	XGBoost + TPE	15.5882	3.9481	**3.0130**	**24.0448**	20.4503	7.7376	0.9030	0.9070	**0.9460**
	LightGBM + TPE	23.7815	4.8766	3.8368	36.4374	25.2577	9.5566	0.8551	0.8582	0.9028
	CatBoost + TPE	**15.3342**	**3.9159**	3.0768	27.0862	**20.2111**	**7.6471**	**0.9045**	**0.9085**	0.9357

*Bold numbers indicate the best result.

models. These results substantiate the effectiveness of hyperparameter optimization using TPE in boosting the predictive capabilities of the EL-models. The improvements in accuracy and generalization achieved by the optimized EL-models signify their potential for practical application in WDN management and decision-making processes.

5.3. Selection of the best EL-model

As discussed in the preceding section, the optimized EL-models have demonstrated satisfactory predictive capabilities. However, determining the overall best EL-model requires a systematic approach, as different EL-models excel in specific evaluation metrics. For instance, when considering the testing dataset, the optimized CatBoost model performed the best in terms of MSE, while the optimized XGBoost model exhibited the highest value for KGE. Moreover, the MAPE values of the optimized RF and XGBoost models are nearly identical. As such, a method that systematically weighs all evaluation metrics is essential to select the best-performing EL-models.

To achieve an unbiased selection for the best EL-model, Copeland algorithm was run on the testing dataset results shown in Table 6, while the results are reported in Table 7, where the model with the highest Copeland points is ranked first. According to the result, the optimized CatBoost model obtained 26 Copeland points, four wins, and zero losses, suggesting its consistent outperformance compared to other EL-models in the majority of the evaluation metrics. The optimized XGBoost model is ranked second with three wins and one loss, indicating its strong performance, although slightly behind the CatBoost model. With two wins and two losses, the RF model garnered a Copeland score of zero, signifying that it achieved an even performance compared to other EL-models. The DT and LightGBM models were ranked fourth and fifth, respectively, with negative Copeland scores, indicating lower performance than the other EL-models. In addition, the Taylor Diagram (TD) is computed using the testing dataset to further confirm the ranking of the EL-models. TD shows three statistical measures, including the standard deviation of the predicted and measured values, the root mean square difference (RMSD), and the correlation coefficient [16]. The standard deviation ratio is represented as the radial distance from the reference point to each model's data point. A shorter radial distance signifies a smaller standard deviation ratio, indicating that the model's variability closely matches the measured data. The RMSD is depicted as the azimuthal angle of each model's data point relative to the reference point. A smaller azimuthal angle indicates a lower RMSD, suggesting better agreement between the model's predictions and the actual measurements. The correlation coefficient is represented by the distance of each model's data point from the reference point. Closeness to the reference point implies a higher correlation between the model's predictions and the measured values. Upon analysis of Fig. 9, it can be observed that the optimized CatBoost model is the closest to the reference point, followed by the optimized XGBoost, RF, DT, and LightGBM models, respectively. This result aligns with the findings obtained from the Copeland algorithm, reinforcing the credibility and consistency of the EL-model ranking.

5.4. Interpretability of the best EL-model

One of the gaps in the previous related studies is that the contribution of the input variables to the model prediction is unknown. To address this limitation, the contribution of the features (i.e., input variables) to the best EL-model predictions is explained using the SHAP framework. Fig. 10 presents the contribution of each feature to the prediction of the optimized CatBoost model. As per the result, "age" has the highest contribution to the predictive EL-model, followed by "installation year,“ "diameter," and "length."

The age and installation year of water pipes are crucial factors in predicting wall-thickness loss. Older pipes are more likely to experience corrosion and wear over time, leading to a decrease in the wall-thickness [61]. Corrosion is a natural process that occurs as pipes are exposed to water and other elements, causing the gradual thinning of the pipe-walls. As pipes age, their resistance to environmental stressors diminishes, making them more vulnerable to damage and leaks. Therefore, age is an essential feature as it captures the cumulative effect of degradation over the pipe lifespan. Regarding the pipe diameter,

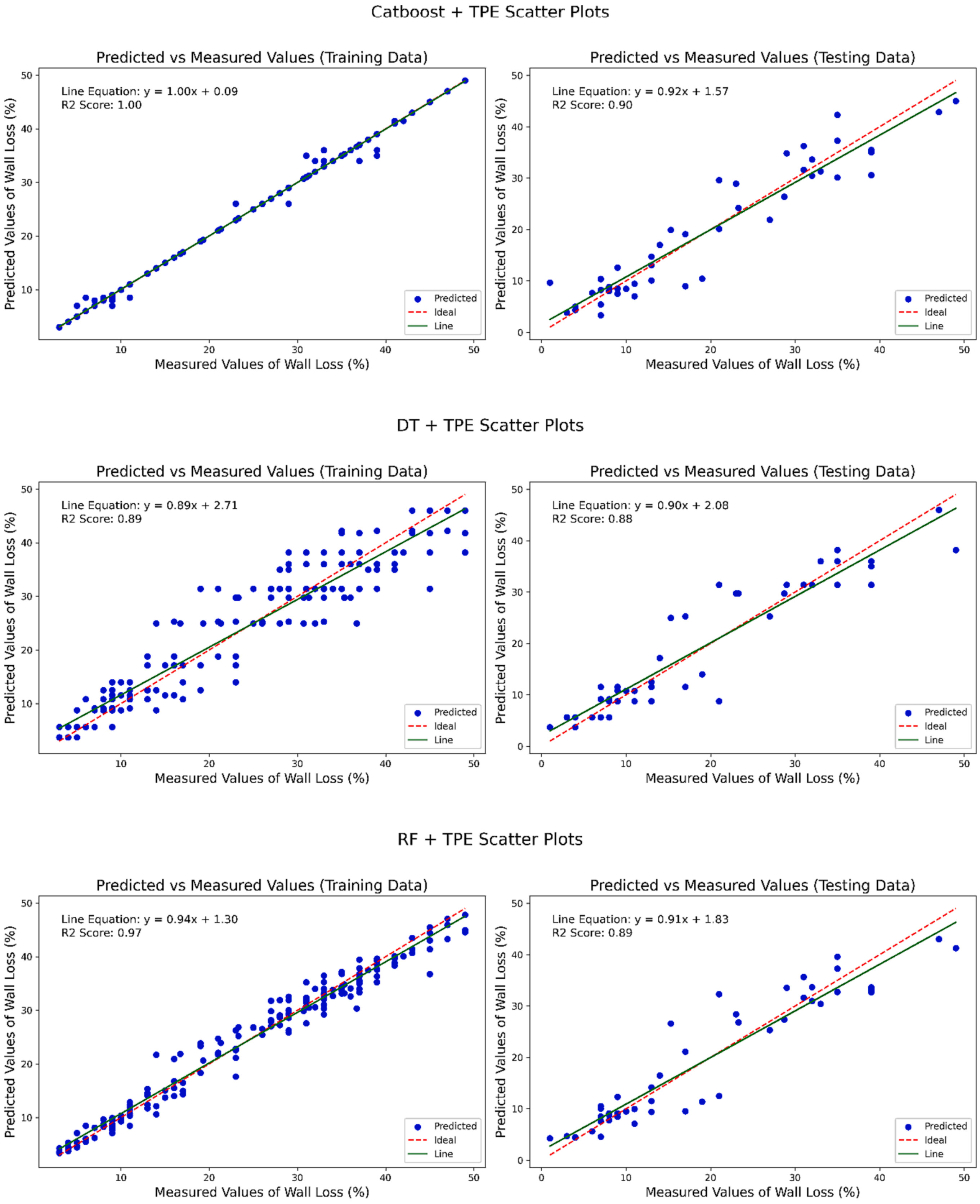

Fig. 8. Scatter plots of the predicted versus measured values of wall-thickness loss for all the five EL-models.

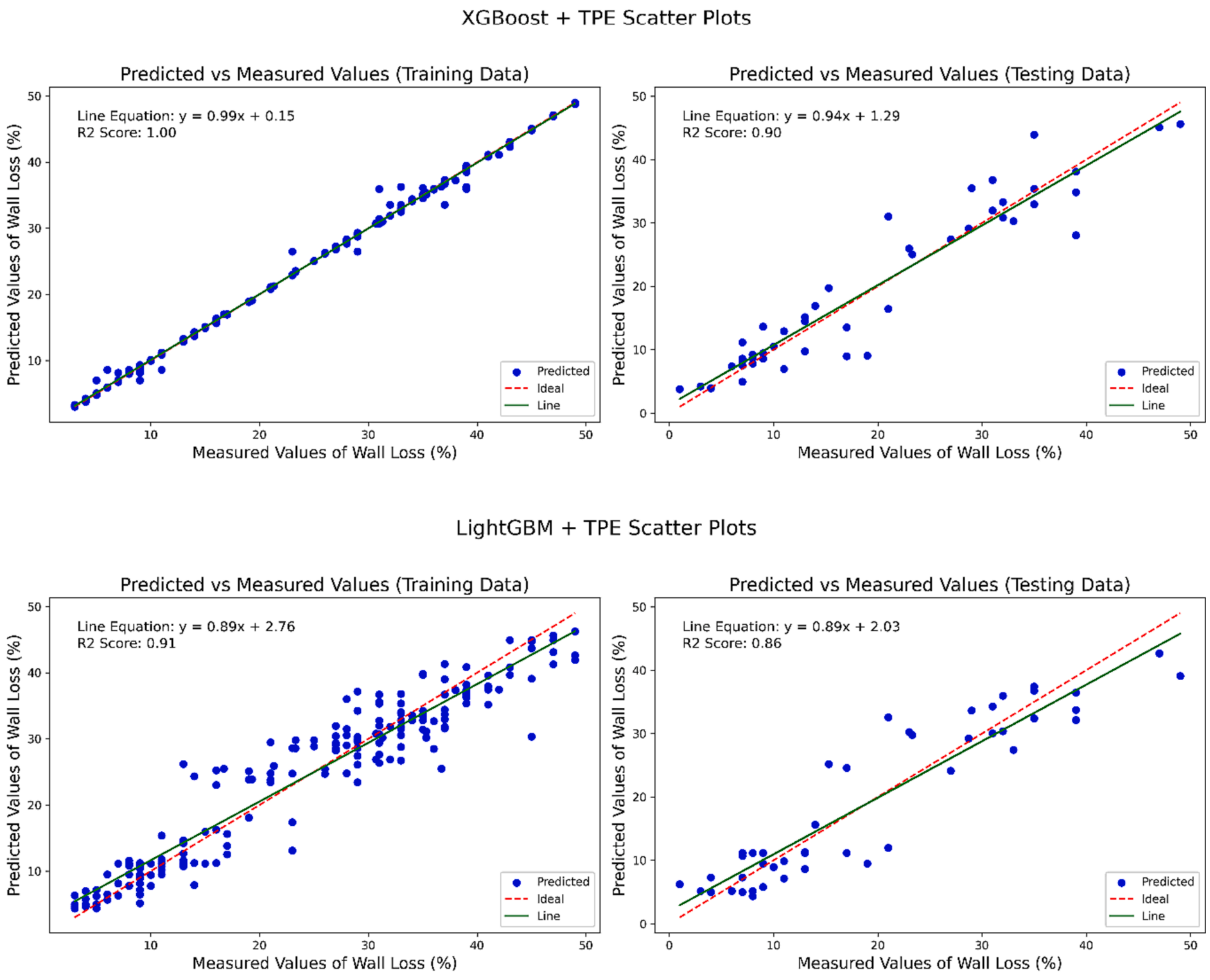

Fig. 8. (*continued*).

Table 7
Comparative result of the Copeland algorithm.

Model	Copeland Point	Wins	Losses	Copeland Score	Rank
CatBoost + TPE	26	4	0	4	1
XGBoost + TPE	24	3	1	2	2
RF + TPE	−2	2	2	0	3
DT + TPE	−12	1	3	−2	4
LightGBM + TPE	−36	0	4	−4	5

variations in water pressure within the pipe can contribute to wall thickness loss. Higher water pressures exert more force on the pipe's internal diameter, leading to potential weakening and thinning of the wall [62,63]. Additionally, pressure variations create cyclical stress on the pipe, promoting fatigue and material degradation. Length is another factor with a high contribution to the predictive EL-model. Longer pipes have a larger surface area exposed to external factors, such as soil movement, temperature variations, and environmental conditions. These external factors can contribute to wall-thickness loss through external corrosion or wear [64].

Fig. 11 visually represents the direction of each feature contribution to the EL-model's prediction. The figure's spectrum reflects the input variables' feature values, ranging from blue (representing the lowest feature value) to red (representing the highest feature value). It is observed that a low value of age (indicated by the blue color) has a negative impact on the prediction. This implies that as the age of the water pipes decreases, the EL-model predicts a lower likelihood of wall-thickness loss. Older pipes, with higher age, are more prone to corrosion, wear, and degradation, leading to a higher probability of wall thickness loss. Therefore, the model correctly captures the inverse relationship between age and wall-thickness loss. The figure also shows that a higher pipe diameter value negatively contributes to the EL-model's prediction, suggesting that larger pipe diameters are associated with a reduced likelihood of wall-thickness loss [65]. Larger diameter pipes can better withstand the stresses and pressures imposed during water transportation, reducing the risk of internal erosion and corrosion that lead to wall-thickness loss.

6. Implications and limitations of the study

6.1. Implications

The optimized EL-models, particularly the CatBoost + TPE and XGBoost + TPE models, demonstrated improved predictive accuracy compared to their base counterparts. These models exhibited better performance across various evaluation metrics, including MSE, RMSE, MAE, and MAPE. Such enhanced predictive accuracy allows water utility managers to make informed decisions about pipe maintenance, replacement, and resource allocation, thereby optimizing the WDN efficiency. The systematic use of the Copeland algorithm allowed for an unbiased selection of the best-performing model. The algorithm considered all evaluation metrics, providing a comprehensive assessment of EL-model performance. The CatBoost + TPE model emerged as the top-performing EL-model based on its Copeland score, with consistent wins across various evaluation metrics. This model selection process ensures that water utility managers adopt the most reliable and accurate

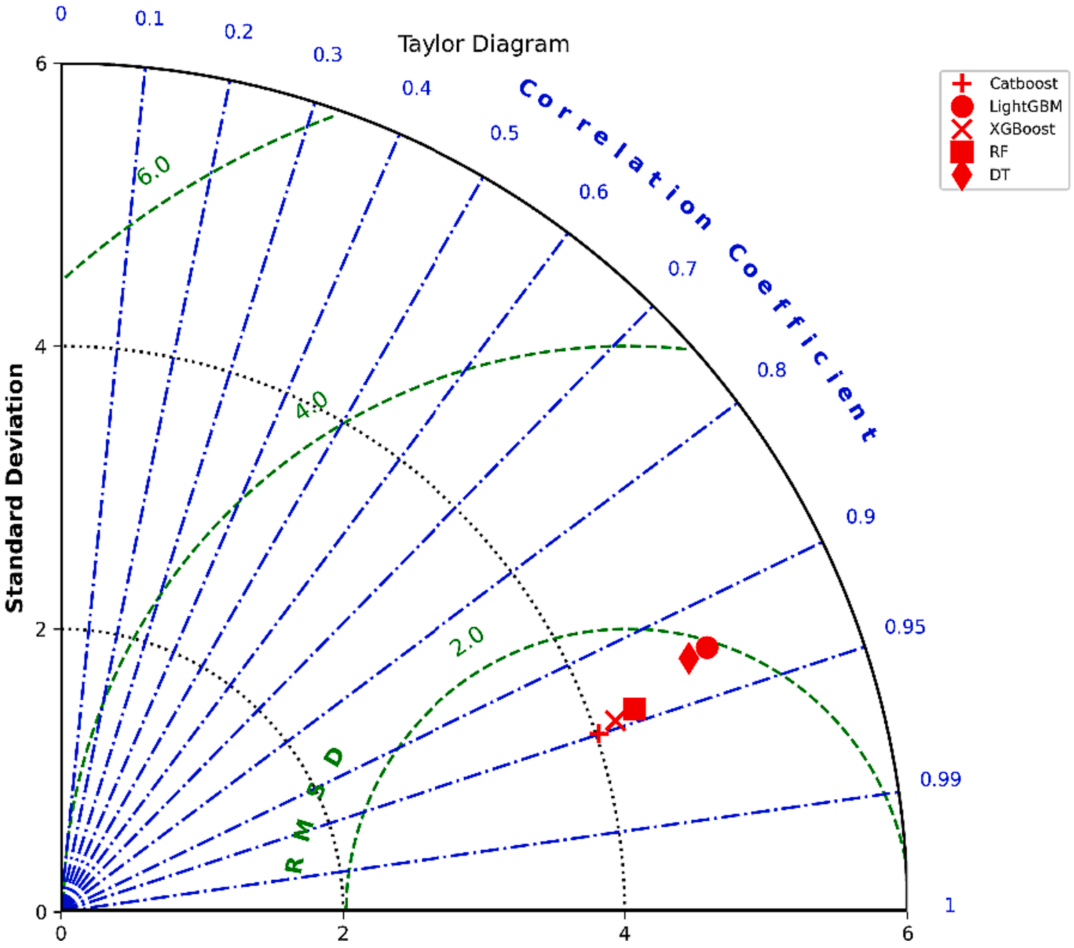

Fig. 9. Taylor diagram of the EL-models using the testing datasets.

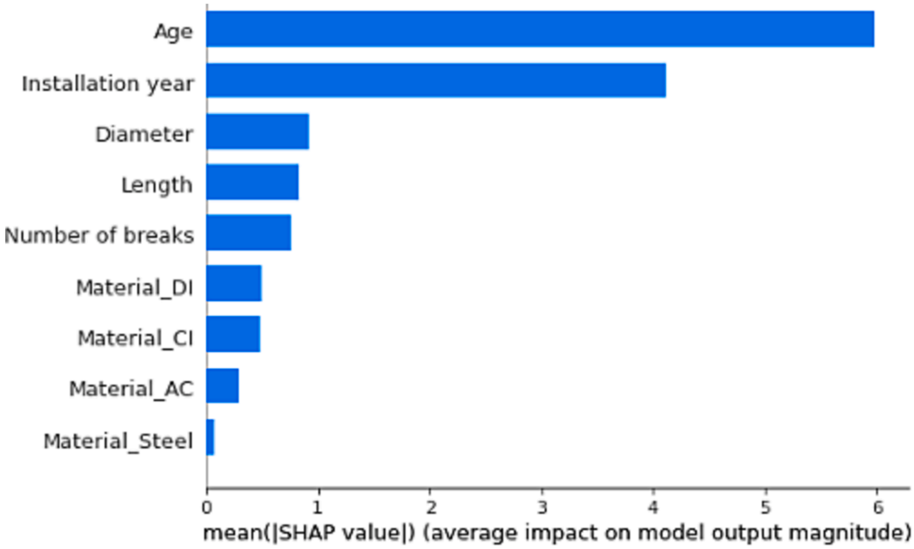

Fig. 10. Contribution of each feature to the prediction of the optimized CatBoost model.

predictive model for their specific WDNs.

The study addressed the lack of model interpretability in previous related studies by employing the SHAP approach to explain the contribution of input variables to the optimized CatBoost model's predictions. The SHAP visualizations provided valuable insights into the influence of age, installation year, diameter, and length on wall thickness loss. This interpretability empowers water utility managers to identify critical factors affecting pipe deterioration and prioritize maintenance efforts effectively. The successful implementation of explainable EL-models with interpretability tools creates a robust decision support system for WDN management. By integrating ML predictions with the SHAP framework, water utility managers comprehensively understand the

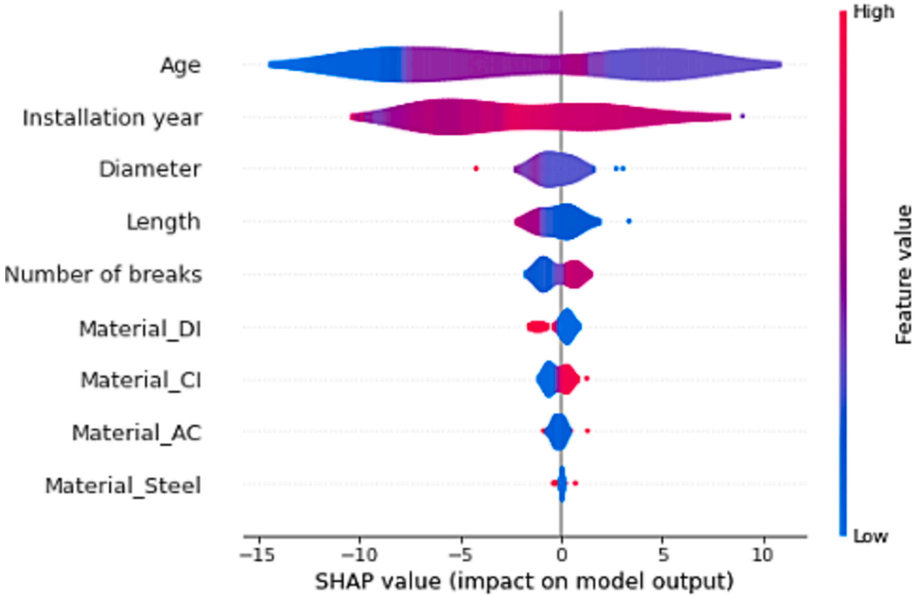

Fig. 11. Directional distribution of the SHAP values on the best EL-model's output.

underlying factors influencing wall-thickness loss. This knowledge facilitates proactive and data-driven decision-making, resulting in cost-effective and sustainable maintenance strategies.

Furthermore, with the ability to predict wall-thickness loss accurately, water utility managers can assess the structural integrity and remaining lifespan of water pipes. This knowledge empowers them to develop targeted rehabilitation plans, prevent catastrophic failures, and enhance the overall WDN resilience. By addressing potential issues before they escalate, the study's implications contribute to the long-term reliability and safety of WDS.

6.2. Limitations

The present study offers valuable contributions to the field of predictive modeling for pipe failure and wall-thickness loss. However, it is crucial to acknowledge the limitations inherent in the research to provide a balanced perspective and guide for future works. One limitation is the relatively small number of factors included in the dataset, consisting of only seven variables, including the dependent variable. This limited set of variables may not fully capture all the complex factors that could influence the wall-thickness loss in water pipes. As a result, there might be some unaccounted variables that could have significant effects on the prediction accuracy of the EL-models.

Additionally, the data used in this study were gathered from limited locations. This geographical restriction may limit the generalizability of the findings to other regions with different environmental conditions and infrastructural characteristics. The dataset may not fully represent the diverse range of conditions and pipe materials found in WDNs globally, thus potentially affecting the models' performance when applied to different contexts. Future research could address these limitations by considering a more comprehensive dataset, including more variables, incorporating data from diverse geographical locations, and exploring alternative modeling approaches.

Another limitation is the absence of a treatment of uncertainty in the predictive models due to lack of data. Incorporating uncertainty considerations into predictive models offers an additional layer of rigor, enhancing the model's applicability for risk assessments and decision-making processes. Uncertainty manifests in various forms, such as model uncertainty, parameter uncertainty, and external factors, as explained below:

- **Model Uncertainty**: This refers to the limitations in the predictive model itself. No model can capture all the complexities of a real-world system, so there's always some level of approximation involved in the model's output.
- **Parameter Uncertainty**: The estimates for model parameters are based on available data and can vary within a confidence interval. This variability introduces another layer of uncertainty.
- **Measurement Uncertainty**: The data employed in this study were measured and recorded using specific instruments and methodologies. Any error or limitation in these processes can contribute to the overall uncertainty of the model predictions.
- **External Factors**: Unaccounted factors such as future environmental changes or shifts in usage patterns can also contribute to prediction uncertainty.

Addressing these limitations could be the focus of future research. For instance, incorporating Bayesian methods or Monte Carlo simulations could offer a more nuanced understanding of uncertainty. This would not only enhance the model's reliability but also provide decision-makers with a more robust basis for risk assessment.

7. Conclusion

Water distribution networks often face significant challenges due to unexpected pipe failures, adversely impacting the environment, economy, and society. To mitigate such failures, one essential approach is predicting the wall-thickness loss of water pipes, an indicator of pipe integrity that has received relatively lower attention in previous studies. Addressing this critical gap, the current research developed and optimized machine learning (ML) models for wall-thickness loss prediction while employing the SHapley Additive exPlanations (SHAP) for model interpretability.

The study harnessed experimental data collected from four WDNs situated in Canada and the USA, encompassing seven variables: age, length, diameter, installation year, material, number of breaks, and wall-thickness loss. Preprocessing steps involved outlier removal and exploratory data analysis, including correlation matrix assessments and other statistical techniques were conducted. For model development, five ensemble algorithms, namely CatBoost, Decision Tree, Random Forest, XGBoost, and LightGBM, were employed to predict wall-thickness loss. Subsequently, the hyperparameters of the EL-models were optimized using the Tree-Structured Parzen Estimator. The EL-models were evaluated using dissimilarity and similarity-based metrics. As per the evaluation metrics, improvements in predictive performance after optimization were demonstrated, affirming the efficiency of hyperparameter tuning in enhancing the models' capabilities. For instance, the MSE and SI of DT using the testing dataset were reduced by 31.46 % and 17.19 %, respectively, after optimizing the hyperparameters. To ensure an unbiased selection of the best optimized model, the Copeland algorithm was employed to systematically rank the models based on the evaluation metrics. The Copeland algorithm's result indicated that CatBoost outperformed the other models, followed by XGBoost and RF. This result was further corroborated by the Taylor Diagram analysis, which displayed the superiority of CatBoost and XGBoost in terms of standard deviation ratio, RMSD, and correlation coefficient. Furthermore, the best model was explained using the SHAP framework. The interpretation of the best model revealed valuable insights into the contribution of input variables to the prediction of wall-thickness loss. The identification of influential features, such as age, installation year, diameter, and length, enables a better understanding of the factors driving pipe degradation, further informing decision-making in WDN management strategies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge the support from the Innovation and Technology Fund (Innovation and Technology Support Programme (ITSP)) under grant number ITS/033/20FP, and the Water Supplies Department of the Hong Kong Special Administrative Region. The statements made herein are solely the responsibility of the authors, and the supporting agencies do not carry any responsibilities or liabilities related to the above findings. The corresponding author acknowledges funding from the European Union's Horizon 2021 research and innovation programme under the Marie Sklodowska-Curie project No-101061320 and the funded APC by Oslo Metropolitan University.

References

[1] Fares H, Zayed T. Hierarchical fuzzy expert system for risk of failure of water mains. J Pipeline Syst Eng Pract 2010;1(1):53–62. https://doi.org/10.1061/(asce)ps.1949-1204.0000037.

[2] Orts E. "America's neglected water systems face a reckoning," *Knowledge at Wharton*, 2015. [Online]. Available: https://knowledge.wharton.upenn.edu/article/americas-neglected-water-systems-face-a-reckoning/#:~:text=Less obvious but at least as,and New York City at 14.2%25.&text=Less obvious but at, York City at 14.2%25.&text=but at least as,and New York City. Accessed: Jul. 10, 2023.

[3] Xu X, Liu S, Smith K, Cui Y, Wang Z. An overview on corrosion of iron and steel components in reclaimed water supply systems and the mechanisms involved. J Clean Prod 2020;276:124079. https://doi.org/10.1016/j.jclepro.2020.124079.

[4] Taiwo R, Ben Seghier MEA, Zayed T. Toward sustainable water infrastructure : the state-of-the- art for modeling the failure probability of water pipes water resources research. Water Resour Res 2023;59. https://doi.org/10.1029/2022WR033256.

[5] Bakhtawar B, Zayed T. State-of-the-art review of leak diagnostic experiments: Toward a smart water network. Wiley Interdiscip. Rev Water 2023;no. April:1–26. https://doi.org/10.1002/wat2.1667.

[6] Taiwo R, Shaban IA, Zayed T. Development of sustainable water infrastructure: A proper understanding of water pipe failure. J Clean Prod 2023;398. https://doi.org/10.1016/j.jclepro.2023.136653.

[7] Poston B., Stevens M. L.A'.s aging water pipes; a $1-billion dilemma. Los Angeles times 2014. [Online]. Available: https://graphics.latimes.com/la-aging-water-infrastructure. Accessed: May 15, 2023.

[8] Paradkar AB. An Evaluation Of Failure Modes For Cast Iron And Ductile Iron Water Pipes (Master's thesis). University of Texas Arlington; 2012.

[9] Centers for Disease Control and Prevention, "Flint Water Crises," 2016. [Online]. Available: https://www.cdc.gov/nceh/casper/docs/FLINT-H.pdf#:~:text=The switch caused water distribution pipes to corrode,NSF International approved filter certified to remove lead. Accessed: March 10, 2023.

[10] Dawood T, Elwakil E, Novoa HM, Delgado JFG. Ensemble intelligent systems for predicting water network condition index. Sustain Cities Soc 2021;73(January). https://doi.org/10.1016/j.scs.2021.103104.

[11] Robles-velasco A, et al. Prediction of pipe failures in water supply networks using logistic regression and support vector classification. Reliab Eng Syst Saf, 196, no. March 2019, p. 106754, 2020, doi: 10.1016/j.ress.2019.106754.

[12] Taiwo R, Ben Seghier MEA, Zayed T. Predicting wall thickness loss in water pipes using machine learning techniques. 2nd Conf Eur Assoc Qual Control Bridg Struct - EUROSTRUCT2023 2023;6(5):1087–92. https://doi.org/10.1002/cepa.2075.

[13] Al-Barqawi H, Zayed T. Infrastructure management: integrated AHP/ANN model to evaluate municipal water mains' performance. J Infrastruct Syst 2008;14(4): 305–18. https://doi.org/10.1061/(asce)1076-0342(2008)14:4(305).

[14] Farh HMH, Ben Seghier MEA, Zayed T. A comprehensive review of corrosion protection and control techniques for metallic pipelines. Eng Fail Anal 2022;vol. 143, no. PA:106885. https://doi.org/10.1016/j.engfailanal.2022.106885.

[15] Ben Seghier MEA, Mustaffa Z, Zayed T. Reliability assessment of subsea pipelines under the effect of spanning load and corrosion degradation. J Nat Gas Sci Eng, 102, no. November 2021, p. 104569, 2022, doi: 10.1016/j.jngse.2022.104569.

[16] Abyani M, Bahaari MR, Zarrin M, Nasseri M. Predicting failure pressure of the corroded offshore pipelines using an efficient finite element based algorithm and machine learning techniques. Ocean Eng, 254, no. November 2021, p. 111382, 2022, doi: 10.1016/j.oceaneng.2022.111382.

[17] Ben Seghier MEA, Plevris V, Solorzano G. Random forest-based algorithms for accurate evaluation of ultimate bending capacity of steel tubes. Structures 2022;44 (April):261–73. https://doi.org/10.1016/j.istruc.2022.08.007.

[18] Amiri-Ardakani Y, Najafzadeh M. Pipe break rate assessment while considering physical and operational factors: a methodology based on global positioning system and data-driven techniques. Water Resour Manag 2021;no. 0123456789. https://doi.org/10.1007/s11269-021-02911-6.

[19] Kutyłowska M. Neural network approach for failure rate prediction. Eng Fail Anal 2015;47:41–8. https://doi.org/10.1016/j.engfailanal.2014.10.007.

[20] Kimutai E, Betrie G, Brander R, Sadiq R, Tesfamariam S. Comparison of statistical models for predicting pipe failures: illustrative example with the city of calgary water main failure. J Pipeline Syst Eng Pract 2015;6(4):04015005. https://doi.org/10.1061/(asce)ps.1949-1204.0000196.

[21] Sattar AMA, Faruk Ö, Gharabaghi B. Extreme learning machine model for water network management. Neural Comput Appl 2017. https://doi.org/10.1007/s00521-017-2987-7.

[22] Opila MC, Attoh-Okine N. Novel approach in pipe condition scoring. J Pipeline Syst Eng Pract 2011;2(3):82–90. https://doi.org/10.1061/(asce)ps.1949-1204.0000081.

[23] Harvey R, McBean EA, Gharabaghi B. Predicting the timing of water main failure using artificial neural networks. J Water Resour Plan Manag 2014;140(4):425–34. https://doi.org/10.1061/(asce)wr.1943-5452.0000354.

[24] Snider B, McBean EA. Improving time-To-failure predictions for water distribution systems using gradient boosting algorithm. 1st Int. WDSA / CCWI 2018 Jt. Conf., no. July, 2018.

[25] Zangenehmadar Z, Moselhi O, Ph D, Eng P. Assessment of remaining useful life of pipelines using different artificial neural networks models. J Perform Constr Facil 2016;30(2007):1–7. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000886.

[26] Fahmy M, Moselhi O. Forecasting the remaining useful life of cast iron water mains. J Perform Constr Facil 2009;23(4):269–75.

[27] Snider B, McBean EA. Combining machine learning and survival statistics to predict remaining service life of watermains. J Infrastruct Syst 2021;27(3):1–14. https://doi.org/10.1061/(asce)is.1943-555x.0000629.

[28] Robles-velasco A, et al. Prediction of pipe failures in water supply networks using logistic regression and support vector classification. Reliab Eng Syst Saf, 196, no. October 2019, p. 106754, 2020, doi: 10.1016/j.ress.2019.106754.

[29] Chen T-Y-J, Vladeanu G, Yazdekhasti S, Daly CM. Performance evaluation of pipe break machine learning models using datasets from multiple utilities. J Infrastruct Syst 2022;28(2):pp. https://doi.org/10.1061/(asce)is.1943-555x.0000683.

[30] Rifaai TM, Abokifa AA, Sela L. Integrated approach for pipe failure prediction and condition scoring in water infrastructure systems. Reliab Eng Syst Saf, 220, no. December 2021, p. 108271, 2022, doi: 10.1016/j.ress.2021.108271.

[31] Geem ZW, Tseng CL, Kim J, Bae C. Trenchless water pipe condition assessment using artificial neural network. Pipelines 2007 Adv. Exp. with Trenchless Pipeline Proj. - Proc. ASCE Int. Conf. Pipeline Eng. Constr., p. 26, 2007, doi: 10.1061/40934 (252)26.

[32] Fan X, Wang X, Zhang X, Xiong PEFA, Yu B. Machine learning based water pipe failure prediction : The effects of engineering , geology , climate and socio-economic factors. Reliab Eng Syst Saf, 219, no. November 2021, p. 108185, 2022, doi: 10.1016/j.ress.2021.108185.

[33] Tavakoli R, Sharifara A, Najafi M. Artificial neural networks and adaptive neuro-fuzzy models to predict remaining useful life of water pipelines razieh. In *World Environmental and Water Resources Congress 2020: ASCE*, 2020, no. 2001, pp. III–IV.

[34] Tavakoli R. Remaining Useful life prediction of water pipes using artificial neural network and adaptive neuro fuzzy inference system models. University of Texas at Arlington; 2018.

[35] Ben Seghier MEA, Keshtegar B, Taleb-Berrouane M, Abbassi R, Trung NT. Advanced intelligence frameworks for predicting maximum pitting corrosion depth in oil and gas pipelines. Process Saf Environ Prot 2021;147:818–33. https://doi.org/10.1016/j.psep.2021.01.008.

[36] Taiwo R, Hussein M, Zayed T. An integrated approach of simulation and regression analysis for assessing productivity in modular integrated construction projects. Buildings 2022;12. https://doi.org/10.3390/buildings12112018.

[37] Cakiroglu C, Shahjalal M, Islam K, Mahmood SMF, Billah AHMM, Nehdi ML. Explainable ensemble learning data-driven modeling of mechanical properties of fiber-reinforced rubberized recycled aggregate concrete. J Build Eng 2023;vol. 76, no. June:107279. https://doi.org/10.1016/j.jobe.2023.107279.

[38] Schober P, Schwarte L. Correlation coefficients: Appropriate use and interpretation. Anesth Analg 2018;126(5):1763–8. https://doi.org/10.1213/ANE.0000000000002864.

[39] Breiman L, Friedman JH, Olshen RA, Stone C. Classification and regression trees. Biometrics 1984;40(3).

[40] Yussif AM, Sadeghi H, Zayed T. Application of machine learning for leak localization in water supply networks. Buildings 2023;13(4):1–21. https://doi.org/10.3390/buildings13040849.

[41] Nowozin S. Improved information gain estimates for decision tree induction. Proc. 29th Int. Conf. Mach. Learn. ICML 2012, vol. 1, pp. 297–304, 2012.

[42] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[43] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat 2001;29(5):1189–232. https://doi.org/10.1214/aos/1013203451.

[44] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-August-2016, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[45] Al Daoud E. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset. Int J Comput Inf Eng 2019;13(1):6–10.

[46] Chen C, Zhang Q, Ma Q, Yu B. LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. Chemom Intell Lab Syst 2019;191(June):54–64. https://doi.org/10.1016/j.chemolab.2019.06.003.

[47] Ke G, et al. LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst, 2017-December, no. Nips, pp. 3147–3155, 2017.

[48] Cakiroglu C, Islam K, Bekdaş G, Nehdi ML. Data-driven ensemble learning approach for optimal design of cantilever soldier pile retaining walls. Structures 2023;51(March):1268–80. https://doi.org/10.1016/j.istruc.2023.03.109.

[49] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. Catboost: Unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst., vol. 2018-December, no. Section 4, pp. 6638–6648, 2018.

[50] Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. J Big Data 2020;7(1):pp. https://doi.org/10.1186/s40537-020-00369-8.

[51] Cakiroglu C, Demir S, Ozdemir MH, Aylak BL, Sariisik G, Abualigah L. Data-driven interpretable ensemble learning methods for the prediction of wind turbine power incorporating SHAP analysis. Expert Syst Appl 2024;237(March):1–12. https://doi.org/10.1016/j.eswa.2023.121464.

[52] Bergstra J, Yamins D, Cox DD. Making a science of model search. pp. 1–11, 2012, [Online]. Available: http://arxiv.org/abs/1209.5111.

[53] Lima LL, Ferreira Junior JR, Oliveira MC. Toward classifying small lung nodules with hyperparameter optimization of convolutional neural networks. Comput Intell 2021;37(4):1599–618. https://doi.org/10.1111/coin.12350.

[54] Chen C, Seo H. Prediction of rock mass class ahead of TBM excavation face by ML and DL algorithms with Bayesian TPE optimization and SHAP feature analysis. Acta Geotech 2023;0123456789:3825–48. https://doi.org/10.1007/s11440-022-01779-z.

[55] Wu J, Chen XY, Zhang H, Xiong LD, Lei H, Deng SH. Hyperparameter optimization for machine learning models based on Bayesian optimization. J Electron Sci Technol 2019;17(1):26–40. https://doi.org/10.11989/JEST.1674-862X.80904120.

[56] Ben Seghier MEA, Keshtegar B, Tee KF, Zayed T, Abbassi R, Trung NT. Prediction of maximum pitting corrosion depth in oil and gas pipelines. Eng Fail Anal 2020;vol. 112, no. March:104505. https://doi.org/10.1016/j.engfailanal.2020.104505.

[57] Taiwo R, Zayed T, Ben Seghier MEA. Integrated intelligent models for predicting water pipe failure probability. Alexandria Eng J 2024;86:243–57. https://doi.org/10.1016/j.aej.2023.11.047.

[58] Furxhi I, Murphy F, Mullins M, Poland CA. Machine learning prediction of nanoparticle in vitro toxicity: A comparative study of classifiers and ensemble-classifiers using the Copeland Index. Toxicol Lett 2019;312(May):157–66. https://doi.org/10.1016/j.toxlet.2019.05.016.

[59] Lundberg S, Lee S-I. A unified approach to interpreting model predictions. In *31st Conference on Neural Information Processing Systems*, 2017, no. 4, pp. 552–564, doi: 10.1016/j.ophtha.2018.11.016.

[60] Fryer D, Strumke I, Nguyen H. Shapley values for feature selection: The good, the bad, and the axioms. IEEE Access 2021;3:3–10. https://doi.org/10.1109/ACCESS.2021.3119110.

[61] Shaban IA, Eltoukhy AEE, Zayed T. Systematic and scientometric analyses of predictors for modelling water pipes deterioration. Autom Constr, 149, no. December 2022, p. 104710, 2023, doi: 10.1016/j.autcon.2022.104710.

[62] Hekmati N, Rahman MM, Gorjian N, Rameezdeen R, Chow CWK. Relationship between environmental factors and water pipe failure: an open access data study. SN Appl Sci 2020;2(11):pp. https://doi.org/10.1007/s42452-020-03581-6.

[63] Zywiec J, Piegdoń I, Tchórzewska-Cieślak B. Failure analysis of the water supply network in the aspect of climate changes on the example of the central and eastern europe region. Sustain 2019;11(24):pp. https://doi.org/10.3390/su11246886.

[64] Zamenian H, Faust KM, Mannering FL, Abraham DM, Iseley T. Empirical assessment of unobserved heterogeneity and polyvinyl chloride pipe failures in water distribution systems. J Perform Constr Facil 2017;31(5):04017073. https://doi.org/10.1061/(asce)cf.1943-5509.0001067.

[65] Wilson D, Filion Y, Moore I. State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. Urban Water J 2017;14(2):173–84. https://doi.org/10.1080/1573062X.2015.1080848.