Journal of Environmental Management 379 (2025) 124738

Contents lists available at ScienceDirect

Journal of Environmental Management

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

Research article

Explainable deep learning models for predicting water pipe failures

Ridwan Taiwo [a,b,*], Tarek Zayed [a], Beenish Bakhtawar [a], Bryan T. Adey [b]

[a] Department of Building and Real Estate, the Hong Kong Polytechnic University, Hung Hom, Hong Kong
[b] Institute of Construction and Infrastructure Management, ETH Zurich, Stefano-Franscini-Platz 5, Zurich, Switzerland

ARTICLE INFO

Handling editor: Jason Michael Evans

Keywords:
Water pipe failure
Probability of leak
Probability of burst
Deep learning
CNN
TabNet
SHAP
Copeland algorithm
Water distribution network

ABSTRACT

Failures within water distribution networks (WDNs) lead to significant environmental and economic impacts. While existing research has established various predictive models for pipe failures, there remains a lack of studies focusing on the probability of leaks and bursts. Addressing this gap, the present study introduces a new approach that harnesses deep learning algorithms — Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and TabNet for failure prediction. The study enhances these base models by optimising their hyperparameters using Bayesian Optimisation (BO) and further refining the models through data scaling. The Copeland algorithm and SHapley Additive exPlanations (SHAP) are also applied for model ranking and interpretation, respectively. Applying this methodology to Hong Kong's WDN data, the study evaluates the models' predictive performance across several metrics, including accuracy, precision, recall, F1 score, Matthews Correlation Coefficient (MCC), and Cohen's Kappa. Results demonstrate that BO significantly enhances the models' predictive abilities, such that the TabNet model's F1 score for leak prediction increases by 36.2% on standardised data. The Copeland algorithm identifies CNN as the most effective model for predicting both leak and burst probabilities. As indicated by SHAP values, critical features influencing model predictions include pipe diameter, material, and age. The optimised CNN model has been deployed as user-friendly web applications for predicting the probability of leaks and bursts, enabling both single-pipe and batch predictions. This research provides crucial insights for WDN management, equipping water utilities with sophisticated tools to forecast the probability of pipe failure, enabling more effective mitigation of such failures.

1. Introduction

1.1. Background

Water distribution networks (WDNs) are critical infrastructures providing essential water resources for residential, commercial, and industrial use (Kerwin and Adey, 2021). Over time, the water pipes that make up these networks can deteriorate and fail, such as through leaks and bursts. These pipe failures can disrupt water delivery and require costly repairs. These pipe failures are caused by a complex interplay of pipe-related, environmental, and operational factors (Al-barqawi & Zayed, 2018; Taiwo et al., 2023a). Understanding the factors contributing to pipe failures is crucial because the average age of WDNs in many countries is increasing, which increases the probability of failure. For example, in the United States, the average age of pipes is over 50 years old, with some systems over 100 years old (Folkman, 2018). As pipes age, failures due to corrosion and fatigue, as well as environmental stresses such as ground movement, temperature fluctuations, and soil corrosivity, are increasingly likely. Less than optimal operational practices, such as irregular pressure changes and lack of preventative maintenance, can further increase the probability of failure over time (Jiang et al., 2019; Wilson et al., 2017).

The consequences of water pipe failures can be categorised into economic, environmental, and social consequences (Li et al., 2024; Mian et al., 2023). Deteriorating water infrastructure leads to a significant amount of non-revenue water - water that is lost before it reaches the end-users due to leaks, bursts, and other failures, representing a direct loss of revenue for water utilities (Kabir et al., 2015). For instance, the USA needs to spend about $1 trillion to rehabilitate its WDNs over the next decades (Fan et al., 2022). In the same vein, China spent about 10 billion RMB in 2014 for pipe replacement and rehabilitation (Xu et al., 2020). The expenditures associated with pipe failure encompass not only the direct costs of repairing the damaged infrastructure but also the additional costs of restoring the affected areas to their original state. Furthermore, pipe failures lead to negative environmental impacts in multiple ways. On average, more than 30% of water in WDNs is lost

[*] Corresponding author. Department of Building and Real Estate, the Hong Kong Polytechnic University, Hung Hom, Hong Kong.
E-mail address: ridwan-a.taiwo@connect.polyu.hk (R. Taiwo).

https://doi.org/10.1016/j.jenvman.2025.124738
Received 24 May 2024; Received in revised form 27 January 2025; Accepted 25 February 2025
Available online 6 March 2025

globally (Tariq et al., 2021). Leaking pipes lose significant volumes of treated drinking water, contributing to water waste and aquifer depletion issues (Tariq et al., 2022; Wasim et al., 2018). Additionally, pipe breaks can cause soil erosion and property damage due to water flooding. Iron corrosion products from ageing pipes also accumulate in soil and groundwater (Taiwo et al., 2023b). Chlorinated water from distribution systems can contaminate nearby surface waters when exfiltration occurs, harming aquatic ecosystems (Liu et al., 2016). Regarding the social consequences, the public experiences many direct inconveniences from pipe failures. Water outages disrupt residential and business activities, while flooded properties cause displacement expenses (Taiwo et al., 2023b). Water pipe failures can pose significant public health risks due to potential water contamination. When cracks, harmful pathogens, and groundwater or soil chemicals can infiltrate the drinking water supply. This contaminated water, if consumed, can then spread waterborne illnesses through local communities. According to the Centers for Disease Control and Prevention (2023), about 7.8 million became sick in the USA due to waterborne diseases.

Based on the aforementioned consequences, it is evident that water pipe failure is a critical issue that needs optimum attention. Predictive modelling is one promising approach to mitigate this serious problem. Traditional prioritisation of pipe replacement relies on risk (combination of failure probability and consequence) assessment methods to prioritise high-risk pipes for rehabilitation (Kerwin and Adey, 2020). These methods include the analytic hierarchy process, index scoring, bow-tie models, fault tree analysis, and fuzzy theory models (Ismaeel and Zayed, 2018; Karamouz et al., 2012; Taiwo et al., 2023b). However, traditional risk assessment has several limitations. First, methods like the analytic hierarchy process depend heavily on subjective human judgment rather than data-driven objectivity (Yeung et al., 2020). Second, traditional risk assessment methods are limited by their labour-intensive nature, making them impractical for large-scale WDNs. Machine learning (ML) has emerged as a powerful alternative, offering automated prediction capabilities that focus on failure probability using historical data patterns (Fan et al., 2022; Giraldo-González and Rodríguez, 2020; Weeraddana et al., 2021). These ML models can efficiently analyse entire networks with minimal manual intervention, enabling rapid and scalable predictions to support proactive maintenance decisions.

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

Reference	Technique	Failure indicator	Evaluation metrics	Type of pipes	Data splitting	Study location
Fan et al. (2022)	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
Rifaai et al. (2022)	LR	Probability of failure	AUC – 0.680 Recall – 0.672 Acc – 0.800	AC, CI, DI, PVC, and others	75% Testing – 25%	Austin, USA
Chen et al. (2022)	XGBoost, RF, B.T.	Probability of failure	AUC = 0.8992	CI, DI, PVC, and others	Training – 12 years data Testing –3 years data	USA
Amiri-Ardakani and Najafzadeh (2021)	MARS, GEP, and M5 Tree	Failure rate	R = 0.981 RMSE = 0.544	AC, CI, PE	Training – 80% Testing – 20%	Yazd, Iran
Snider and McBean (2021)	WPHSM, RF, and RSF	Remaining useful life	C-Index = 0.925	AC, CI, and DI	Training – 80% Testing – 20%	Canada
Dawood et al. (2021)	ANFIS and FIS	Condition index	R^2 = 0.9145 RMSE = 0.6829	–	Training – 60% Testing – 40%	Arequipa, Peru
Giraldo-González and Rodríguez (2020)	GBT, Bayes, SVM, and ANN	Probability of failure	Acc – 0.9979 F1 score – 0.4643	AC ad PVC	Training – 70% Testing – 30%	Bogota, Colombia
Kerwin et al. (2020)	ANN	Time to failure	R = 0.882	CI, DI, and PE	Training – 80% Testing – 20%	Switzerland
Robles-velasco et al. (2020)	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
Tavakoli et al. (2020)	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
Snider and McBean (2018)	ANN, RF, and XGBoost	Time to failure	R – 0.85 RMSE – 5.81	AC, CI, DI, and PVC	Training – 80% Testing – 20%	North America
Winkler et al. (2018)	DT, RF, AdaBoost, RUSBoost	Probability of failure	AUC – 0.93	AC, CI, DI, Steel, PE, and PVC	Training – 50% Testing – 50%	Austria
Sattar et al. (2017)	Extreme Learning Machine	Failure rate	R^2 – 0.65 RMSE – 0.09	AC, CI, and DI	Training – 75% Testing – 25%	Toronto, Canada
Zangenehmadar et al. (2016)	ANN	Remaining useful life	R^2 – 0.9877 MAE – 3.890 MAPE – 2.870	AC, CI, Concrete, DI, PE, PVC, Steel, and Copper	Training – 70% Testing – 30%	Quebec, Canada
Kutyłowska (2015)	ANN	Failure rate	R^2 – 0.4142	CI, PE, PVC, Steel	Training – 75% Testing – 25%	Poland
Kimutai et al. (2015)	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
Harvey et al. (2014)	ANN	Time to failure	R – 0.82 RE – 0.32	AC, CI, DI, and PVC	Training – 70% Testing – 30%	Scarborough, Canada
Fahmy and Moselhi (2009)	MLP, GRNN, and MR.	Remaining useful life	R^2 = 0.96 MAE = 0.12	CI	Training – 80% Testing – 20%	USA and Canada
Geem et al. (2007)	ANN	Condition index	R^2 = 0.8629	CI, DI, and Steel	–	South Korea

1.2. Previous studies and research gaps

Researchers have adopted ML algorithms, including linear regression, logistic regression, random forest (RF), gradient boosting machine (GBM), support vector machine (SVM), extreme learning machine (ELM), adaptive network-based fuzzy inference system (ANFIS), amongst others (Dawood et al., 2021; Robles-velasco et al., 2020; Snider and McBean, 2018) to predict failures of water pipes. The failure indicators that have been predicted using these models include the probability of failure, time-to-failure, condition index, remaining service life, and failure rate (Taiwo et al., 2023a). Table 1 presents a summary of ML studies predicting water pipe failure, showcasing the techniques employed, indicators of failure predicted, evaluation metrics adopted, types of pipes examined, data division ratio, and the geographical location of the study. In instances where a study has developed multiple models, Table 1 reports the evaluation metrics of the best-performing model.

Snider and McBean (2021) developed a random survival forest (RSF) model that outperformed conventional modelling approaches using historical data of a utility in Canada. By accounting for right-censored data on pipes still in service, the RSF model could estimate failure rate more accurately than standalone random forest or parametric Weibull models. It was found that the RSF approach could reduce pipe replacement and repair costs by 14% over a 50-year planning horizon for the water utility. Furthermore, Kutyłowska (2015) used 11 years of historical data of a WDN (distribution and house connection pipes) in Poland to develop an ANN model for predicting pipe failure rates. Multilayer perceptron (MLP) was employed as the type of ANN, while the quasi-Newton algorithm was used as the learning method. According to the result, the R^2 for house connections was 0.95, while for distribution pipes, it was 0.92 for the training dataset. The R^2 dropped to 0.41 for the house connection pipes while that of the distribution pipes was negative.

Four models (Gradient-Boosted Tree, Bayes, SVM, and ANN) were used to predict the probability of failure for individual pipes, using physical, environmental, and operational variables as inputs (Giraldo-González and Rodríguez, 2020). The models were trained on data covering 61,251 pipes in an 1819 km Colombian WDN. GBT performs better than the other models, giving more importance to the misclassified pipes. The GBT model automatically identifies the most influential variables through its iterative training process. For this network, the number of previous failures, pipe length, and precipitation were the most important factors in predicting failures. The results showed that around 0.17% of the pipes have a high probability of failure in the present condition, which requires appropriate maintenance or replacement strategies. Similarly, Winkler et al. (2018) compared the performance of different decision tree methods, such as simple decision trees, RF, AdaBoost, and RUSBoost, using a real-world case study of a medium-sized city in Austria. RUSBoost outperformed the other algorithms, such that it had an AUC value of 0.93.

Sattar et al. (2017) developed an extreme learning model (ELM) model to predict the time-to-failure of pipes in a WDN located in the Greater Toronto Area (Canada). The model outperformed other ML algorithms, such as ANN, non-linear regression, and SVM. The authors employed Monte Carlo simulation and differential evolution algorithm to estimate the uncertainty and sensitivity of the ELM model predictions. It was discovered that the number of previous pipe breaks is the most influential input parameter, followed by pipe diameter. They also found that cathodic protection (CP) is more effective than cement mortar lining (CML) in increasing failure time for ductile iron (DI) pipes and cast iron (CI). Furthermore, an ANN model was developed to predict the time-to-failure of individual pipes in a WDN located in Ontario, Canada. The network spans 5850 km, consisting of 6346 pipes. A separate model was developed for each pipe material, including AC, CI, and DI. The models' effectiveness was assessed using relative error and correlation coefficients on testing and holdout datasets. It was determined that the impact of prior failure emerged as the most crucial variable across all models.

In order to predict the aggregated condition index of WDNs in Peru, an ensemble framework that combines ANFIS with the fuzzy inference system (FIS) was developed (Dawood et al., 2021). The framework consists of two modules, the first one using ANFIS to predict the condition index of each WDN based on six parameters, and the second module consolidating the provinces' indices into one regional index using FIS. The performance of the proposed framework was validated by comparing it with a multiple linear regression (MLR) model. Based on the four evaluation metrics: R-square, adjusted R-square, sum of squares due to error, and root mean squared error, it was demonstrated that the proposed framework outperformed the MLR model in all metrics. Similarly, Tavakoli et al. (2020) employed seven input variables (material, age, length, diameter, installation year, number of previous failures, and wall thickness) to predict the remaining useful life of water pipes in the USA and Canada. ANN and ANFIS were adopted for the modelling, which showed comparative results. The study deduced that, on average, around 10% loss in wall thickness in existing CI, DI, AC, and steel pipes leads to a reduction of about 50% in the remaining useful life.

Although ML algorithms, including simple ANN architecture, have been widely used for predicting different failure indicators of water pipes, the potential of deep learning (DL) algorithms is yet to be fully explored in this domain. DL is a subset of ML that employs more complex multilayer neural network architectures that can learn deeper representations and patterns from data. DL models have potential advantages over shallow neural networks and other traditional ML techniques for pipe failure modelling. These models are highly scalable and can learn from large datasets with thousands or millions of pipes to improve generalizability (Raziani and Azimbagirad, 2022). Innovations in DL, like attention layers and convolution layers, can learn spatial and temporal dependencies related to pipe failure, which shallow ML cannot capture (al-Ani et al., 2023). Existin g water pipe failure prediction models have focused on predicting the likelihood of failure events occurring without distinguishing between failure types like leaks versus bursts (Rifaai et al., 2022; Robles-velasco et al., 2020). However, leaks and bursts have distinct failure mechanisms and driving factors. Leaks often result from corrosion-induced holes and joints over time, while bursts are sudden ruptures from excessive internal pressure (Pękala and Pietrucha-Urbanik, 2018; Rezaei et al., 2015). Another gap in the extant literature is the lack of a systematic approach for selecting the optimal model when multiple algorithms are developed and compared. Previous studies typically train a variety of models such as ANNs, SVMs, and decision trees on a pipe failure dataset, assess performance metrics like accuracy for each model, and informally select the superior model based on the highest metrics. Nonetheless, this method can exhibit bias as different algorithms may surpass others depending on the metric being evaluated.

Based on the aforementioned gap in the literature, the aim of this study is to fill these gaps and contribute to the existing knowledge. Thus, the objectives of this study are stated below:

- **Development of optimised probability of leak and burst models**: This study focuses on developing optimised models for predicting the probability of leak and burst of water pipes using deep neural network (DNN), convolution neural network (CNN), and TabNet.
- **Selection of the best-optimised model**: This study adopts the Copeland algorithm to compare and rank the optimised DL model to prevent biased model selection.
- **Interpretation of the selected optimised model**: This study employs the SHapley Additive exPlanations (SHAP) framework to interpret the superior model. This framework delineates the contribution of each feature to the predictive model, indicating whether the impact of each feature is either positive or negative.

2. Research methodology

The framework adopted in this study is depicted in Fig. 1, outlining a detailed methodology for developing explainable DL models to predict water pipe failures. The framework is divided into five sequential steps, each contributing to creating a robust and interpretable predictive model. The first step is data preparation. This initial step involves the selection and processing of relevant data. The data is categorised into three types: pipe-related, environment-related, and operation-related. These datasets are then divided into a training and validation set (70% of the data) and a testing set (30%). Key pre-processing tasks such as data cleaning, outlier removal, imputation for missing values, normalisation, and standardisation are applied to ensure the data is suitable for training the DL models. Individual cleaning is performed on training and testing datasets separately to prevent data leakage. The training and validation subset is subjected to a 10-fold cross-validation procedure to optimise model generalizability. In this process, the data is split into ten folds, wherein one fold serves as the validation set while the remaining nine folds comprise the training set. In the second step, DL architectures such as DNN, CNN, and TabNet are considered. Bayesian optimisation is used to tune the hyperparameters of these models. This process involves training surrogate models on a training dataset and using them to predict the performance of the DL models on a validation dataset. The optimisation process iteratively proposes new points (sets of hyperparameters) to find the set that maximises performance on the validation set. In the third stage, the models' performance is assessed using a set of evaluation metrics: Accuracy, Recall, Precision, F1 score, Matthews Correlation Coefficient (MCC), and Cohen's Kappa. These metrics provide a broad overview of the models' predictive capabilities and performance. Subsequently, the Copeland method, a pairwise comparison ranking algorithm, is utilised to rank the models. Each model's performance is compared against the others', with wins, losses, and Copeland scores (the difference between the number of wins and losses) being calculated. This score effectively ranks the models according to their predictive abilities. The final step focuses on providing explainability for the chosen deep learning model using SHAP. The analysis of marginal contributions allows for a better understanding of the model's decision-making process. Additionally, the distribution of SHAP values can be used to identify the impact of the features on the model, which can be either positive or negative. The details of the proposed methodology are explained in subsequent sections.

3. Data collection and pre-processing

Data were acquired from the Water Supplies Department (WSD) of HK, the entity tasked with overseeing the WDN within the region. The WDN spans over 8300 km and is the primary water supply infrastructure for a population exceeding 7.41 million (Water Supplies Department HKSAR, 2021). The data provision from the WSD comprised two GIS files, one detailing the configuration of the water network and the other cataloguing incidences of leaks and bursts across the network from the years 2010–2020. Upon integration of these data, the resulting compilation included records for 1,089,232 pipes. Of these pipes, 37,767 were identified to have suffered from leaks, while bursts were noted in 1552 cases. This indicates that leakages and bursts affected a mere 3.47% and 0.142% of the network, respectively. To address the inherent class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was implemented. SMOTE generates synthetic samples of the minority classes by interpolating between existing minority class samples rather than simple duplication. Importantly, SMOTE was applied only to the training dataset, while the test dataset was maintained in its original state to ensure model evaluation reflects real-world conditions where failure events are naturally rare. This approach ensures that model performance metrics represent realistic operational scenarios while addressing the training challenges posed by imbalanced data.

Additional data were sourced from various open databases to form a more holistic view of the variables influencing pipe failures. This supplementary data encompasses climatic variables obtained from the HK Observatory and traffic information sourced from the HK Transport Department. Subsequent to the dataset aggregation, a total of 14 variables were identified and categorised into three groups: those pertaining to the pipes themselves, environmental conditions, and operational characteristics. The group of pipe-specific variables encompassed attributes such as length, diameter, material, and service age. Environmental variables included the corrosiveness of the soil, the type of roadway above, the surrounding land's utilisation, meteorological factors such as temperature, precipitation, and humidity, and the annual average daily traffic (AADT) values. The operational variables consisted of the water pressure within the pipes, water type, and whether CP was implemented or not. The dataset's descriptive statistics are presented in Table 2. A binary system was employed for categorical variables, assigning a '1′ to denote the presence of a characteristic in a given data point and a '0′ to signify its absence (i.e., one-hot encoding).

Pre-processing of data is a critical stage in the development of DL models, as the accuracy of predictions is greatly influenced by the quality of the input data. To address issues of redundancy, noise, and heterogeneity in the data, several pre-processing techniques were employed. Initial steps included identifying and removing outliers, utilising methods such as box and scatter plots, and a thorough examination of descriptive statistics. For instance, pipe data indicating an age of 115 years, which represented a scant proportion of less than 1%, were deemed outliers and thus removed. For the treatment of missing entries, numerical data points were substituted with the mean of their respective feature, while the most frequent values, or modes, were used for categorical data. The rationale behind these methods is to maintain the integrity of the dataset's distribution. In the scaling phase, both normalisation and standardisation procedures were executed, drawing on their demonstrated effectiveness in prior research (Almheiri et al., 2021; Robles-velasco et al., 2020). The mathematical expressions for these scaling methods are presented in Equations (1) and (2). Normalisation rescales the data into a fixed range, typically from 0 to 1, based on the maximum and minimum values of each feature. On the other hand, standardisation modifies the data to fit a standard normal distribution, centralising the mean at zero with a standard deviation of one (Uddin et al., 2022).

$$m_{norm} = \frac{m_i - m_{min}}{m_{max} - m_{min}} \tag{1}$$

$$m_{stand} = \frac{m_i - m_{mean}}{m_{std}} \tag{2}$$

where m_i, m_{norm} *and* m_{stand} refer to the unscaled, normalised, and standardised value of a data instance and $m_{min}, m_{max}, m_{mean}$ *and* m_{std} represent the minimum, maximum, mean, and standard deviation of a feature.

4. Model development

4.1. Predictive models using deep learning algorithms

The selection of DNN, CNN, and TabNet architectures for this study was based on several technical considerations. DNN was selected as a baseline deep learning model due to its established capability in processing tabular data and modelling complex non-linear relationships between features (Almheiri et al., 2021). The architecture's feedforward nature makes it particularly suitable for the classification task of pipe failure prediction. CNN was incorporated for its ability to learn spatial hierarchies of features automatically. While traditionally utilised in image processing, CNNs have demonstrated effectiveness in structured tabular data where local patterns may exist between neighbouring features (Akinosho et al., 2020). This capability is particularly relevant for pipe failure prediction, where relationships between physical and

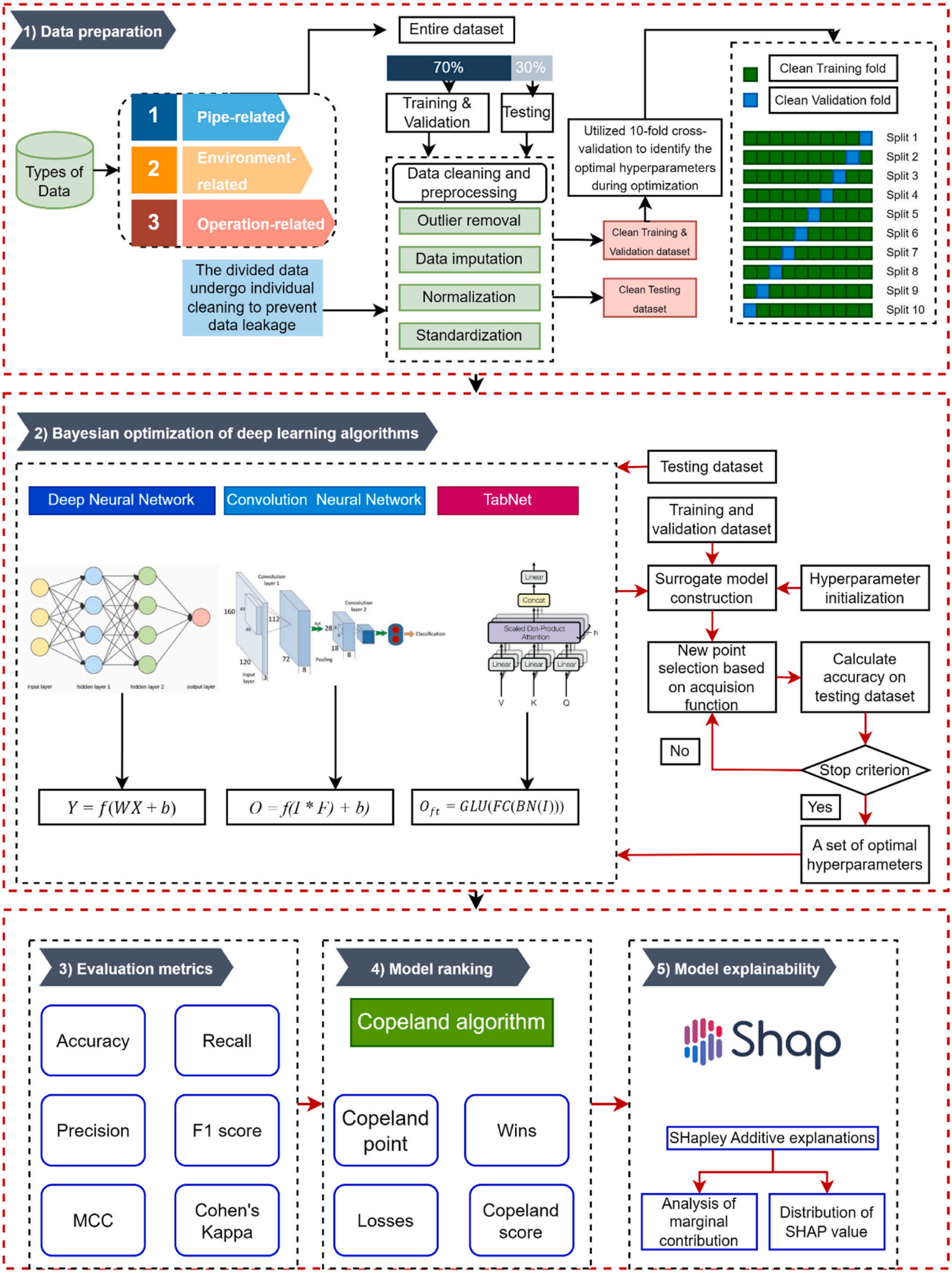


Fig. 1. Framework of the study.

Table 2
Descriptive statistics of the data.

Factor	Unit	Mean	Std	Min	Max
Length	Metre (m)	7.43	17.43	1.00	200.00
Diameter	Millimeters (mm)	131.21	167.44	20.00	3000.00
Material	AC, CI, DI, PE	–	–	0	1
Age	Years	24.17	20.29	0	70.00
Soil corrosivity	Non-corrosive, Mildly corrosive, Highly corrosive	–	–	0	1
Road type	Footway, Carriageway, Other locations	–	–	0	1
Land use	Urban, Rural.	–	–	0	1
Temperature	°C	24.26	6.96	2.30	30.20
Precipitation	Millimeters (mm)	10.56	9.07	1.02	792.00
Traffic	AADT	11,427.88	15,970.03	1.00	179,400.00
Pressure	Bars	6.24	2.42	0.53	24.60
Water type	Freshwater, Saltwater	–	–	0	1
CP	–	–	–	0	1

operational characteristics may exhibit spatial dependencies. TabNet was selected as it represents a state-of-the-art architecture specifically designed for tabular data processing. Its distinctive feature selection mechanism and interpretability through sequential attention layers make it particularly applicable for infrastructure management applications where understanding the model's decision-making process is essential (Arık and Pfister, 2021).

4.1.1. Deep neural network

The Deep Neural Network (DNN) is a foundational model in the domain of DL algorithms (Chen and Xu, 2024; Zheng et al., 2024). DNNs are characterised by their depth, which comprises multiple hidden layers between the input and output layers, enabling them to model complex and high-level abstractions in data (Jun et al., 2017).

A standard DNN architecture is composed of an input layer X, multiple hidden layers H, and an output layer Y. Each layer consists of units or neurons, and each neuron in one layer is connected to every neuron in the subsequent layer, forming a dense network. The input layer receives the feature vectors derived from the data, which are then processed through the hidden layers using a series of weighted summations and non-linear activation functions.

The mathematical operations within a typical hidden layer l can be represented as follows:

$$\boldsymbol{H}^{(l)} = \sigma\left(\boldsymbol{W}^{(l)}\boldsymbol{H}^{(l-1)} + \boldsymbol{b}^{(l)}\right) \tag{1}$$

where $H^{(l-1)}$ is the output of the previous layer or the input data for the first hidden layer, $W^{(l)}$ denotes the weight matrix, $b^{(l)}$ is the bias vector, and σ represents the non-linear activation function, such as ReLU or sigmoid, applied element-wise. This process is iteratively conducted across all hidden layers.

The output layer of DNN provides the final prediction, which is framed as a classification task, with the objective of predicting the probability of leak and burst for water pipes. For such classification tasks, the output layer typically employs an activation function, which outputs a probability distribution across the classes, which is represented mathematically as:

$$\boldsymbol{Y} = \sigma\left(\boldsymbol{W}^{(output)}\boldsymbol{H}^{(last)} + \boldsymbol{b}^{(output)}\right) \tag{2}$$

Here, Y is the vector of probabilities that the pipe section falls into each of the possible outcome classes.

4.1.2. Convolution neural network

The 1D Convolutional Neural Network (CNN) architecture is specifically designed to process the multi-dimensional data associated with the characteristics of water pipes to predict their leakage/burst probabilities. The input layer is structured to receive pre-processed feature matrices representing the pipe-related, environment-related, and operation-related factors. The convolutional layers form the core of the CNN architecture. Multiple convolutional layers are employed to extract and learn features from the input data (Raziani and Azimbagirad, 2022; Tsai et al., 2022). The first convolutional layer applied a set of learnable filters (kernels) $\boldsymbol{W}^{(l)}$, where l represents the layer number. The convolution operation at each layer l for a given input matrix $\boldsymbol{X}^{(l)}$ is defined in Equation (3).

$$\boldsymbol{F}^{(l)} = \boldsymbol{W}^{(l)} * \boldsymbol{X}^{(l)} + \boldsymbol{b}^{(l)} \tag{3}$$

where $\boldsymbol{F}^{(l)}$ is the feature map obtained after applying the kernel $\boldsymbol{W}^{(l)}$ to the input $\boldsymbol{X}^{(l)}$, and $\boldsymbol{b}^{(l)}$ represents the bias.

After each convolution operation, an activation function is applied to introduce non-linearity, enabling the network to learn complex patterns. Following the convolutional layers, pooling layers are used to reduce the spatial size of the representation, decreasing the number of parameters and computations in the network. Max pooling is utilised, which can be defined using Equation (4).

$$\boldsymbol{P}^{(l)}_{i,j} = max_{m,n\in M_{i,j}}\, \boldsymbol{A}^{(l)}_{m,n} \tag{4}$$

where $P^{(l)}$ is the pooled feature map and $M_{i,j}$ is the region in the activated feature map $A^{(l)}$ over which the pooling operation is performed.

The high-level reasoning in the network is performed by fully connected layers. The output from the final pooling layer is flattened and fed into a series of fully connected layers. The operation at a fully connected layer l with the input vector $\boldsymbol{v}^{(l)}$ is given in Equations (5) and (6).

$$z^{(l)} = \boldsymbol{W}^{(l)}\,\boldsymbol{v}^{(l)} + \boldsymbol{b}^{(l)} \tag{5}$$

$$o^{(l)} = \sigma\left(z^{(l)}\right) \tag{6}$$

where $z^{(l)}$ is the linear combination of weights $\boldsymbol{W}^{(l)}$, biases $\boldsymbol{b}^{(l)}$, and input $\boldsymbol{v}^{(l)}$ and $o^{(l)}$ is the output after applying the activation function σ. It should be noted that the final fully connected layer acted as the output layer, with a single neuron using the sigmoid activation function to predict the probability p of pipe leakage or burst, given by Equation (7):

$$p = \frac{1}{1 + e^{z(output)}} \tag{7}$$

where z^{output} is the input to the output neuron

4.1.3. TabNet

TabNet is a novel DL architecture for tabular data that uses sequential attention to learn which features to focus on during the learning process (Arık and Pfister, 2021). Fig. 2 gives a schematic representation of TabNet architecture, which uses an encoder composed of multiple steps to determine relevant features from the data and generates a feature representation (Nguyen and Byeon, 2023). This representation is then aggregated to assist in decision-making. The model's input, consisting of batch-sized data with D-dimensional features, undergoes batch normalisation before being processed by the feature transformer.

The feature transformer is structured with multiple gated linear unit (GLU) blocks. These GLU blocks, which include fully connected and batch normalisation layers, are designed for robust learning, with some blocks being shared and others independent. To maintain stability and control variance, normalisation is applied after each GLU block. The

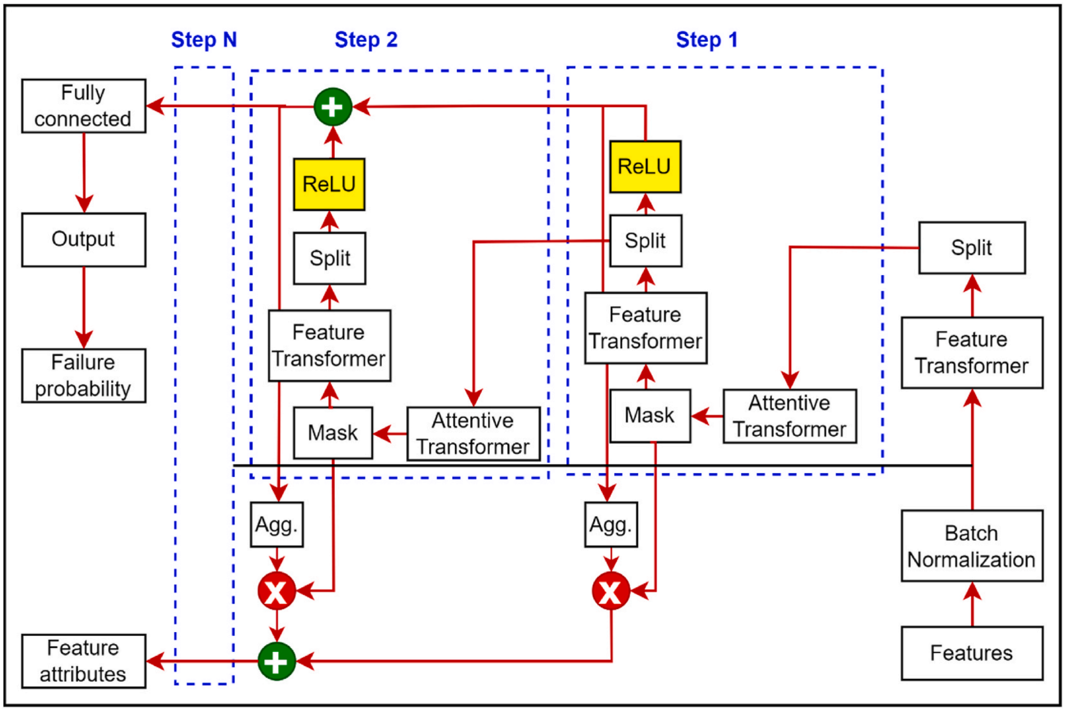


Fig. 2. Schematic representation of the TabNet architecture.

resulting transformed features are then fed into the attentive transformer.

The attentive transformer is composed of fully connected and batch normalisation layers, followed by a prior scale and sparsemax layer. It uses the information from the prior step to calculate a mask layer for the current step, as shown in Equations (8) and (9):

$$P[i] = \prod_{j}^{i} (\gamma - M[j]) \tag{8}$$

$$M[i] = sparsemax\ (P[i-1]^{*}h_i(a[i-1])) \tag{9}$$

where $P[i]$ represents the 'prior scale' at the i-th decision step, γ is the relaxation parameter, the mask from previous layers is represented by $M[j]$ and $h_i(\cdot)$ represents the trainable function of the fully connected and batch normalisation layers. Sparsemax is used to ensure the sum of the mask coefficients is 1, contributing to sparse feature selection.

To control the sparsity of the features, sparsity regularisation L_{sparse} is introduced in the form of entropy to the loss function, adding a small number $\in$ for numerical stability, which is depicted in Equation (10).

$$L_{sparse} = \sum_{i=1}^{N_{steps}} \sum_{b=1}^{B} \sum_{j=1}^{D} M_{bj}[i] \log(M_{bj}[i]) + \in \tag{10}$$

The feature transformer takes the masked features and outputs them for both the decision process and the next attentive transformer step, using Equation (11).

$$[d[i], a[i] = f_i(M[i]^{*}f) \tag{11}$$

where $[d[i]$ is the decision step output, and $a[i]$ is the information for the subsequent step.

Regarding the interpretability, TabNet computes the importance of each step through an aggregation of the output vector, which is converted into a scalar, reflecting the step's significance to the final outcome. Local feature importance for a sample is derived by summing the results across all steps, and global feature importance is calculated using an aggregate mask.

4.2. *Hyperparameter optimisation using BO*

After building the DL models, their hyperparameters are optimised using Bayesian Optimisation (BO), a strategy that enhances model performance by efficiently navigating the hyperparameter space. The hyperparameters and search space optimised for these models are presented in Table 3. BO seeks to identify the optimal set of parameters for a model. The approach aims to find the maximum value of an unknown objective function $\theta(p)$, at a given point p, within a defined search space Ω. The optimal sampling point can be represented by Equation (12) (Bello et al., 2024; Taiwo et al., 2023b; Yang and Shami, 2020).

$$p^{+} = \underset{p \in \Omega}{\operatorname{argmax}} \theta(p) \tag{12}$$

The BO process involves two key steps:

Table 3
Details of the hyperparameter optimisation.

Model	Hyperparameter	Type	Range
DNN	Batch size	Integer	[8, 128]
	Epochs	Integer	[2, 50]
	Number of neurons	Integer	[8, 64]
	Optimizer	Categorical	[Adam, SGD, Adagrad, RMSprop]
	Learning rate	Continuous	[0.01, 1]
CNN	Batch size	Integer	[8, 128]
	Epochs	Integer	[2, 50]
	Number of neurons	Integer	[8, 64]
	Optimizer	Categorical	[Adam, SGD, Adagrad, RMSprop]
	Number of filters	Integer	[8, 128]
	Kernel size	Integer	[2, 20]
	Learning rate	Continuous	[0.01, 1]
TabNet	n_d	Integer	[8, 128]
	n_a	Integer	[8, 128]
	n_steps	Integer	[8, 128]
	gamma	Continous	[1, 10]
	lambda_sparse	Continuous	[0.01, 1]
	Batch size	Integer	[8, 128]
	Step size	Continuous	[0.01, 1]

1. A surrogate probabilistic model is fitted to the objective function, often using a Gaussian Process (GP), which is then updated as new data points are sampled. This model is preferred for its flexibility, robustness, and analytic tractability.
2. An acquisition function is constructed from the posterior distribution of the surrogate model to balance the search space exploration with the exploitation of known good regions. The Expected Improvement (EI) is a common choice for the acquisition function. The optimisation process iterates, continually updating the surrogate model with new findings, until a predefined stopping criterion, typically the maximisation of the acquisition function, is met.

To address overfitting concerns in the DL models, several methodological and architectural safeguards were implemented. The dataset was partitioned into training and validation (70%) and testing (30%) sets, with careful separation to prevent data leakage. A rigorous 10-fold cross-validation procedure was employed during model training. Additionally, specific architectural features were incorporated in each model: the DNN architecture included dropout layers with a rate of 0.2 between dense layers and L2 regularisation on layer weights. The CNN model implemented a combination of dropout, batch normalisation after convolutional layers, and early stopping. TabNet's inherent sparse feature selection mechanism served as a natural regulariser. The hyperparameters of all three models were optimised using Bayesian optimisation, which intelligently searches the hyperparameter space to find configurations that maximise validation set performance while avoiding overfitting.

4.3. Evaluation metrics

The predictive algorithms generate a continuous score ranging from 0 to 1, reflecting the likelihood of a pipe experiencing a leak or burst. Given that the target variable within the dataset is dichotomous, it is standard practice to select a cutoff point for categorising the outcomes (Robles-velasco et al., 2020; Taiwo et al., 2024a). In this study, pipes with predicted failure probabilities above 0.5 are classified by the model as failures, while those below 0.5 are classified as non-failures. A confusion matrix is constructed to compare the model's predictions against the actual conditions of the pipes, as depicted in Fig. 3 (Mazumder et al., 2021; Robles-velasco et al., 2020). This matrix is then used to calculate six different performance metrics, which are detailed in Table 4.

Table 4
Evaluation metrics for the DL models.

Evaluation metric	Mathematical expression
Accuracy	$\frac{TL/B + TI}{TL/B + TI + FL/B + FI}$
Precision	$\frac{TL/B}{(TL/B + FL/B)}$
Recall	$\frac{TL/B}{(TL/B + FI)}$
F1 score	$2^{*}\frac{(Precision^{*}Recall)}{(Precision + Recall)}$
MCC	$\frac{(TL/B \times TI - FL/B \times FI)}{\sqrt{TL/B + FL/B) \times (TL/B + FI) \times (TI + FL/B) \times (TI + FI}\Big)}$
Cohen's Kappa	$\frac{Accuracy - Expected\ accuracy}{1 - Expected\ accuracy}$ Expected accuracy $=$ $\frac{((TL/B + FI) \times (TL/B + FL/B) + ((TI + FL/B) + (TI + FI)}{(TL/B + TI + FL/B + FI)^{2}}$

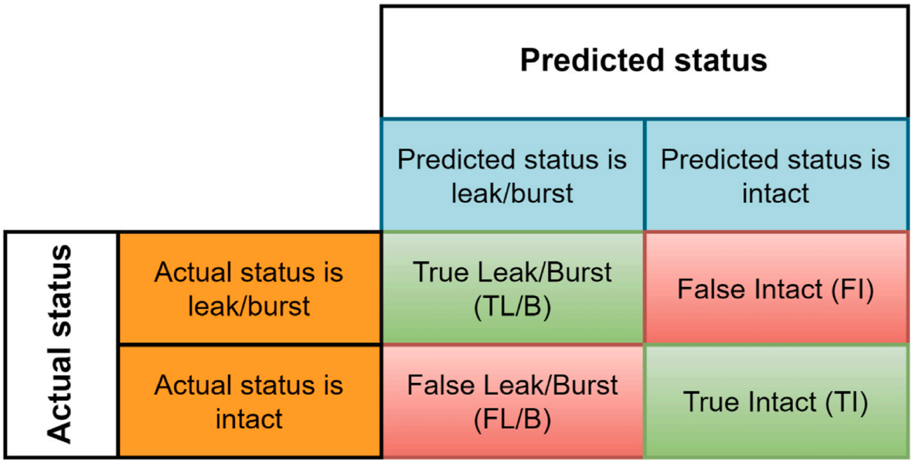
4.4. Ranking of the DL models

The Copeland method, a paired comparison approach, is employed to rank the performance of the DL models. This method involves comparing each model against every other in a series of head-to-head matchups. The process of evaluation is highlighted as follows (Furxhi et al., 2019; Taiwo et al., 2024b):

1. **Pairwise Comparisons**: Each DL model is compared with every other model for each performance metric. In these comparisons, models are awarded points based on their performance relative to one another.
2. **Points Allocation**: A model earns a point for each performance metric where it outperforms another model. For instance, if Model A has a higher precision than Model B in their comparison, Model A receives a point.
3. **Win/Loss Record**: The outcome of each pairwise comparison is a win, loss, or tie for the models involved. A win is recorded for a model if it accrues more points than the other model in their comparison. Similarly, a loss is noted when a model earns fewer points than its competitor, while a tie is considered if both models accumulate an equal number of points.
4. **Copeland Score Calculation**: The Copeland score for each model is calculated by subtracting the total number of losses from the total number of wins:

$$Copeland\ Score = Total\ Wins - Total\ Losses \tag{13}$$

5. **Ranking of Models**: Models are then ranked based on their Copeland scores, with the model boasting the highest score at the top



		Predicted status	
		Predicted status is leak/burst	Predicted status is intact
Actual status	Actual status is leak/burst	True Leak/Burst (TL/B)	False Intact (FI)
	Actual status is intact	False Leak/Burst (FL/B)	True Intact (TI)

Fig. 3. Confusion matrix for classifying water pipe status.

rank, indicating it has the best performance across the evaluated metrics.

4.5. Model interpretability using SHAP

To elucidate the decision-making process of DL models, SHAP values are utilised to interpret the best model selected by the Copeland Method. The contribution of each individual feature to the prediction made by the model relative to a baseline is quantified using SHAP values. The baseline typically represents the average model output over the dataset. For a given prediction, the SHAP value for feature k is formulated using Equation (14).

$$\psi_{k(p)} = \sum\nolimits_{T \subseteq G\{k\}} \frac{|T|!(|G| - |T| - 1)!}{|G|!} \left[h_{T\cup\{k\}}(x_p) - h_T(x_p)\right] \tag{14}$$

where:

- $\psi_{k(p)}$ symbolises the SHAP value for feature k for prediction instance p.
- G denotes the total set of features.
- T is a subset of features not including feature k.
- $|T|$ signifies the number of elements in subset T.
- $|G|$ is the count of all features.
- $h_{T\cup\{k\}}(x_p)$ represents the prediction with the inclusion of feature k.
- $h_T(x_p)$ is the prediction without feature k.

5. Model evaluation and discussion

This section presents and discusses the results of the base-DL models, optimised-dl models, selection of the best DL model, and its interpretability.

5.1. Results of the base-DL models

After training the DL models, their efficiency was evaluated using the testing dataset presented in this section. Table 5, which was generated using the confusion matrix (see Table S1), displays the performance of these base-DL models, highlighting differences among scaling methods and models. Given the dataset's highly imbalanced nature, where leaks are relatively rare compared to non-leak instances, accuracy alone is not a sufficient measure of model performance. This imbalance also contributes to the generally lower precision across all models, as the number of true positive predictions is small relative to the number of false positive predictions. When comparing the scaling methods, it's evident that both normalisation and standardisation improve the performance across all models compared to non-scaled data. This improvement is particularly notable in the precision metric, which is critical in imbalanced datasets. For example, the precision for the DNN model increases from 0.584 with non-scaled data to 0.640 with normalised and 0.726 with standardised data, indicating that the likelihood of correctly identifying a leak has increased with proper scaling. Recall is robust across models, indicating a strong ability to identify actual leaks. The F1 score, which balances precision and recall, is highest with the CNN model in standardised data and is closely followed by CNN, underscoring its efficiency in managing the trade-off between identifying leaks and avoiding false alarms. MCC and Cohen's Kappa provide a more nuanced view of the models' performance by considering true negatives and the imbalance in the dataset. These metrics are particularly high with standardised data, reflecting a better true positive rate relative to the imbalance in the dataset. Similarly, Table 6 presents the evaluation metrics for the probability of burst models (the confusion matrix is presented in Table S2). According to the table, recall values are high for all models, suggesting that the models are generally successful at identifying the majority of the actual burst events. In most cases, the performance of the model increases when the data is normalised or standardised. For instance, the F1 score of TabNet improved by 32.6% when the data was normalised. Generally, the results show that CNN and TabNet outperformed DNN.

5.2. Results of the optimised-dl models

Table 7 presents the evaluation metrics for DL models that have undergone optimisation to predict the probability of leaks in water pipes (see Table S3 for the confusion matrix). Consistent with the base-DL models, the performance metrics of these optimised models demonstrate improvement with data normalisation and standardisation, reinforcing the significance of data scaling in DL applications. Markedly, the precision and recall of the optimised DNN model showed increases of 12.7% and 8.4%, respectively, upon standardisation. While normalisation typically enhances model performance, the optimised TabNet model's slight decrease in precision and F1 score suggests its potential preferential alignment with non-scaled data. The standardised DNN and CNN models exhibit high F1 scores and MCC values, indicative of a well-calibrated balance between precision and recall, alongside a robust alignment between predicted outcomes and actual occurrences. When contrasting the base and optimised DL models, the superior performance of the latter becomes apparent, underscoring the effectiveness of hyperparameter tuning. For instance, the non-scaled DNN model's precision, recall, and F1 score surged by 23.8%, 8.7%, and 21.0%, respectively, post-optimisation.

According to Table 8 (see Table S4 for the confusion matrix), which shows the evaluation metrics for the probability of burst models, TabNet generally outperforms DNN and CNN in non-scaled and normalised datasets, particularly in precision and F1 scores. CNN shows a substantial increase in performance metrics with normalised and standardised data, indicating it may be sensitive to the scaling method applied. DNN shows less variation in performance across scaling methods but doesn't reach the precision value of CNN or TabNet with normalised data. A discerning examination of Tables 7 and 8 indicates that standardised data attained the highest performance for models predicting the probability of leaks, whereas normalised data showed optimal results for models predicting the probability of bursts. This distinction underscores the tailored impact of data scaling techniques on model efficacy, contingent on the specific predictive task at hand. Using the normalised data, Figs. 4 and 5 visualise the difference between the base and optimised DL models for the probability of leak and burst

Table 5
Evaluation metrics for base-DL models for predicting the probability of leaks.

Data scaling	Models	Accuracy	Precision	Recall	F1 score	MCC	Cohen's Kappa
Non-scaled	DNN	0.974	0.584	0.861	0.696	0.696	0.682
	CNN	0.975	0.603	0.866	0.711	0.711	0.698
	TabNet	0.966	0.512	0.842	0.637	0.641	0.62
Normalised	DNN	0.978	0.640	0.882	0.742	0.741	0.731
	CNN	0.981	0.668	0.914	0.772	0.772	0.762
	TabNet	0.974	0.598	0.835	0.697	0.694	0.684
Standardised	DNN	**0.985**	0.726	0.915	0.809	0.807	0.801
	CNN	**0.985**	0.725	**0.924**	**0.812**	**0.811**	**0.805**
	TabNet	**0.985**	**0.727**	0.903	0.805	0.803	0.798

Table 6
Evaluation metrics for base-DL models for predicting the probability of bursts.

Data scaling	Models	Accuracy	Precision	Recall	F1 score	MCC	Cohen's Kappa
Non-scaled	DNN	0.997	0.247	0.914	0.388	0.474	0.387
	CNN	0.997	0.268	0.924	0.415	0.496	0.414
	TabNet	0.997	0.270	0.937	0.419	0.502	0.418
Normalised	DNN	0.998	0.332	0.947	0.492	0.56	0.491
	CNN	0.998	0.366	0.950	0.529	0.589	0.528
	TabNet	0.998	0.393	0.950	0.556	0.61	0.555
Standardised	DNN	0.998	0.388	0.950	0.551	0.607	0.550
	CNN	0.998	**0.433**	0.963	**0.597**	**0.645**	**0.597**
	TabNet	**0.999**	0.395	**0.967**	0.561	0.618	0.561

*numbers in bold represent the best result.

Table 7
Evaluation metrics for optimised-dl models for predicting the probability of leaks.

Data scaling	Models	Accuracy	Precision	Recall	F1 score	MCC	Cohen's Kappa
Non-scaled	DNN	0.988	0.767	0.936	0.843	0.841	0.837
	CNN	0.991	0.826	0.950	0.883	0.881	0.879
	TabNet	0.993	0.865	0.957	0.908	0.906	0.905
Normalised	DNN	0.990	0.796	0.954	0.868	0.867	0.863
	CNN	0.990	0.801	0.946	0.867	0.865	0.862
	TabNet	0.989	0.793	0.942	0.861	0.859	0.856
Standardised	DNN	**0.994**	0.865	0.969	0.914	0.913	0.911
	CNN	**0.994**	**0.876**	**0.978**	**0.924**	**0.923**	**0.922**
	TabNet	0.990	0.799	0.938	0.863	0.861	0.858

Table 8
Evaluation metrics for optimised-dl models for predicting the probability of burst.

Data scaling	Models	Accuracy	Precision	Recall	F1 score	MCC	Cohen's Kappa
Non-scaled	DNN	0.999	0.586	0.973	0.732	0.755	0.731
	CNN	0.997	0.307	0.980	0.468	0.548	0.467
	TabNet	0.999	0.738	0.980	0.842	0.85	0.841
Normalised	DNN	0.999	0.733	0.977	0.838	0.846	0.837
	CNN	0.999	0.782	**0.987**	**0.872**	**0.878**	**0.872**
	TabNet	0.999	**0.783**	0.983	**0.872**	0.877	**0.872**
Standardised	DNN	0.999	0.608	0.983	0.751	0.773	0.751
	CNN	0.999	0.743	0.980	0.845	0.853	0.845
	TabNet	0.999	0.697	0.987	0.817	0.829	0.817

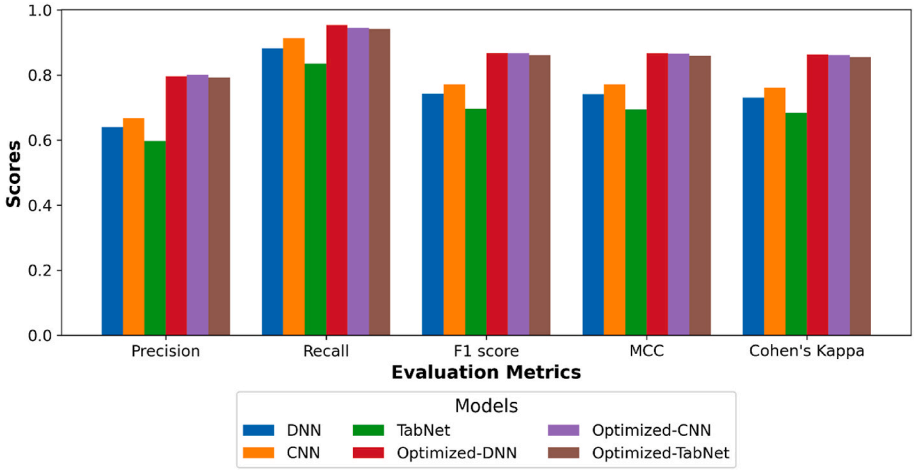


Fig. 4. Comparison of base and optimised-dl models for probability of leak models.

models.

Analysis of the performance metrics across training and testing datasets reveals effective management of overfitting and underfitting in the DL models. The implemented dropout layers and regularisation techniques, combined with the optimisation strategies, helped the models achieve higher performance without overfitting, as evidenced by the balanced improvement in both precision and recall metrics. For leak prediction, the standardised CNN model achieved high F1 scores (0.924) and MCC values (0.923) without compromising generalisation, while for burst prediction, the normalised CNN and TabNet models both achieved

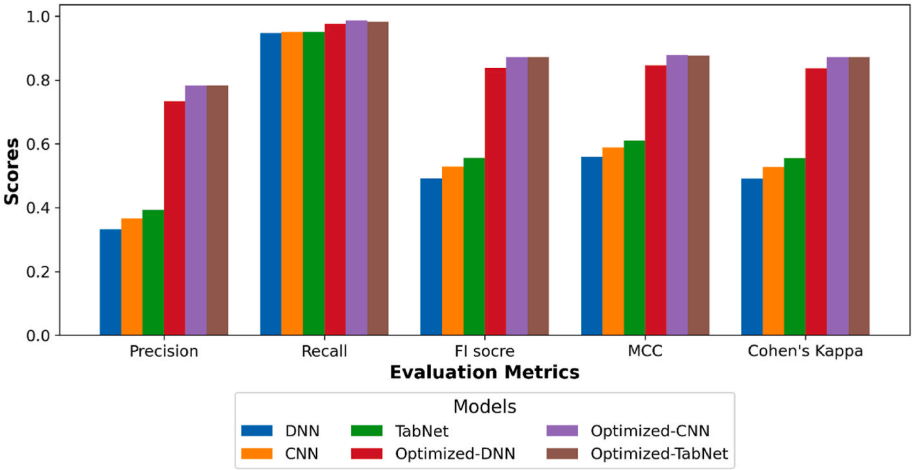


Fig. 5. Comparison of base and optimised-dl models for probability of burst models.

F1 scores of 0.872. The similar performance patterns across different data scaling methods (non-scaled, normalised, and standardised), together with the architectural features like batch normalisation and early stopping, further suggest that the models maintain good generalisation capabilities. The implemented cross-validation procedure and the combination of dropout and regularisation have successfully prevented underfitting, as demonstrated by the high recall values (>0.93 for leak prediction and >0.97 for burst prediction) across all optimised models, indicating their ability to capture the underlying patterns in the data effectively.

5.3. Selection of the best model

The optimised models have shown high performance in predicting the probability of leaks and bursts. Yet, selecting the top-performing model necessitates a structured method since various models show superiority in different evaluation metrics. For example, considering the probability of bursts in normalised datasets, the optimised TabNet surpasses its counterparts in precision, whereas the CNN stands out with the highest recall. Additionally, both CNN and TabNet share equivalent scores in terms of the F1 metric and Cohen's Kappa. Thus, employing a systematic evaluation approach that considers all metrics is crucial to ascertain the most effective model. Tables 9 and 10 show the results of the Copeland algorithm for ranking the optimised probability of leak and burst models.

Since data scaling has proved to improve the models' predictive capability, the comparison is made between normalised and standardised datasets. With normalised data, DNN performed the best with 2 wins, 0 losses, and a Copeland score of +2, giving it rank 1. For standardised data, CNN emerged superior with 2 wins, 0 losses, and a score of +2, putting it in 1st place. Looking at the aggregate scores and ranks across both scaling methods, CNN performed the best overall with 14 Copeland points and a Copeland score of +2, securing it the top rank. DNN took 2nd place with an aggregate record of 10 Copeland points. TabNet came last with 0 wins, 4 losses, and a score of −4. The superior performance of CNN in this context suggests its potential as a reliable choice for predicting the probability of leaks for the WDN.

Table 9
Results of the Copeland algorithm for ranking the probability of leak models.

Data scaling	Model	Copeland Point	Wins	Losses	Copeland Score	Rank
Normalised	DNN	9	2	0	2	1
	CNN	3	1	1	0	2
	TabNet	−12	0	2	−2	3
Standardised	DNN	1	1	1	0	2
	CNN	11	2	0	2	1
	TabNet	−12	0	2	−2	3
Aggregate	DNN	10	3	1	2	2
	CNN	**14**	**3**	**1**	**2**	**1**
	TabNet	0	0	4	−4	3

Table 10
Results of the Copeland algorithm for ranking the probability of burst models.

Data scaling	Model	Copeland Point	Wins	Losses	Copeland Score	Rank
Normalised	DNN	−10	0	2	−2	3
	CNN	6	2	0	2	1
	TabNet	4	1	1	0	2
Standardised	DNN	−8	0	2	−2	3
	CNN	6	2	0	2	1
	TabNet	2	1	1	0	2
Aggregate	DNN	−18	0	4	−4	3
	CNN	**12**	**4**	**0**	**4**	**1**
	TabNet	6	2	2	0	2

Table 10 demonstrates consistency in model performance across both scaling methods. For the normalised data set, CNN takes the lead with the highest Copeland point, score, and rank, suggesting its better adaptability to normalised data, unlike DNN, which falls to the lowest rank. The standardised data set also sees CNN and TabNet performing well, but DNN lags behind. The aggregate scores across both data sets reveal CNN's superiority over others. The overall results highlight the importance of considering different data scaling methods when evaluating model performance. Each model's strengths and weaknesses become apparent under varying conditions, underscoring the necessity of choosing a model not only based on its overall accuracy but also on its adaptability to different data representations.

5.4. Interpretability of the best model

Following the evaluative methodology delineated in Section 5.3, the optimised CNN model utilising standardised data for leak prediction and normalised data for burst prediction has been identified as the most performant. Figs. 6 and 7 illustrate the importance of the feature as determined by SHAP values for models predicting the probability of leak and burst in water pipes, respectively. For the leak prediction model, 'Diameter', 'Material_Plastic', and 'Age' are identified as the leading features influencing the model's predictions. These attributes suggest that the model places significant emphasis on the physical properties and the material composition of the pipes, along with their operational lifespan. In contrast, the burst prediction model prioritises 'Diameter' and 'Material_Plastic', similar to the leak model, but assigns greater

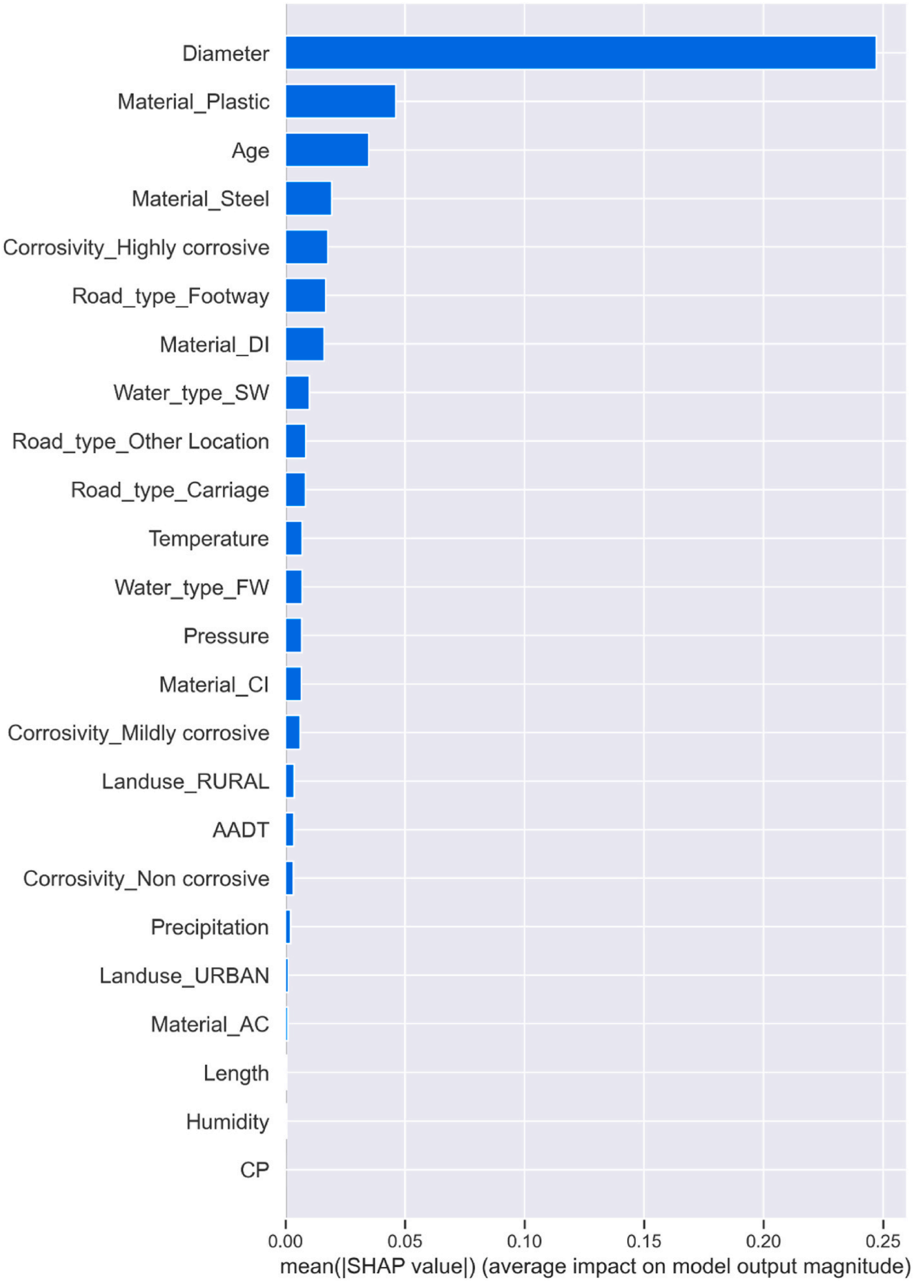


Fig. 6. SHAP feature importance for the probability of leak model.

importance to 'Corrosivity_Highly corrosive' conditions. This differentiation in feature importance highlights the distinct mechanisms and factors the model associates with the likelihood of burst incidents as opposed to leaks.

Figs. 8 and 9 elucidate the relationship between feature values and their SHAP values, which quantify the impact on the model's predictions. The colour gradient in these figures, transitioning from red to blue, represents the spectrum of feature values across the dataset, with red signifying higher values and blue indicating lower ones. For continuous variables, this gradient reflects a range of values, while for categorical variables, represented as dummy variables, the visualisation simplifies to red and blue, denoting the presence or absence of the feature, respectively.

According to these figures, it is observed that smaller diameters contribute positively to the model's prediction of an event (leak or burst), as indicated by positive SHAP values (Taiwo et al., 2023a). Conversely, greater ages are similarly associated with positive SHAP values, suggesting that the probability of an event increases with the age of the pipe. These insights are consistent with extant literature and domain knowledge. While SHAP values may indicate correlative relationships, the analysis reveals causal mechanisms underlying these correlations (Dillon et al., 2018). For instance, the high importance of 'Diameter' in both leak and burst predictions can be explained through established mechanical principles: smaller diameter pipes have higher surface area-to-volume ratios, making them more susceptible to environmental factors, and experience greater pressure fluctuations due to their lower flow capacity (Barton et al., 2019; Taiwo et al., 2023c). Additionally, smaller diameters are more prone to blockages due to reduced flow areas, leading to increased localised pressures and subsequent failures. The significance of 'Material_Plastic' in both models reflects specific material degradation mechanisms. Plastic pipes, while resistant to chemical corrosion, are susceptible to UV degradation, thermal stress, and mechanical damage. Over time, these factors cause molecular chain scission and oxidation, leading to reduced material

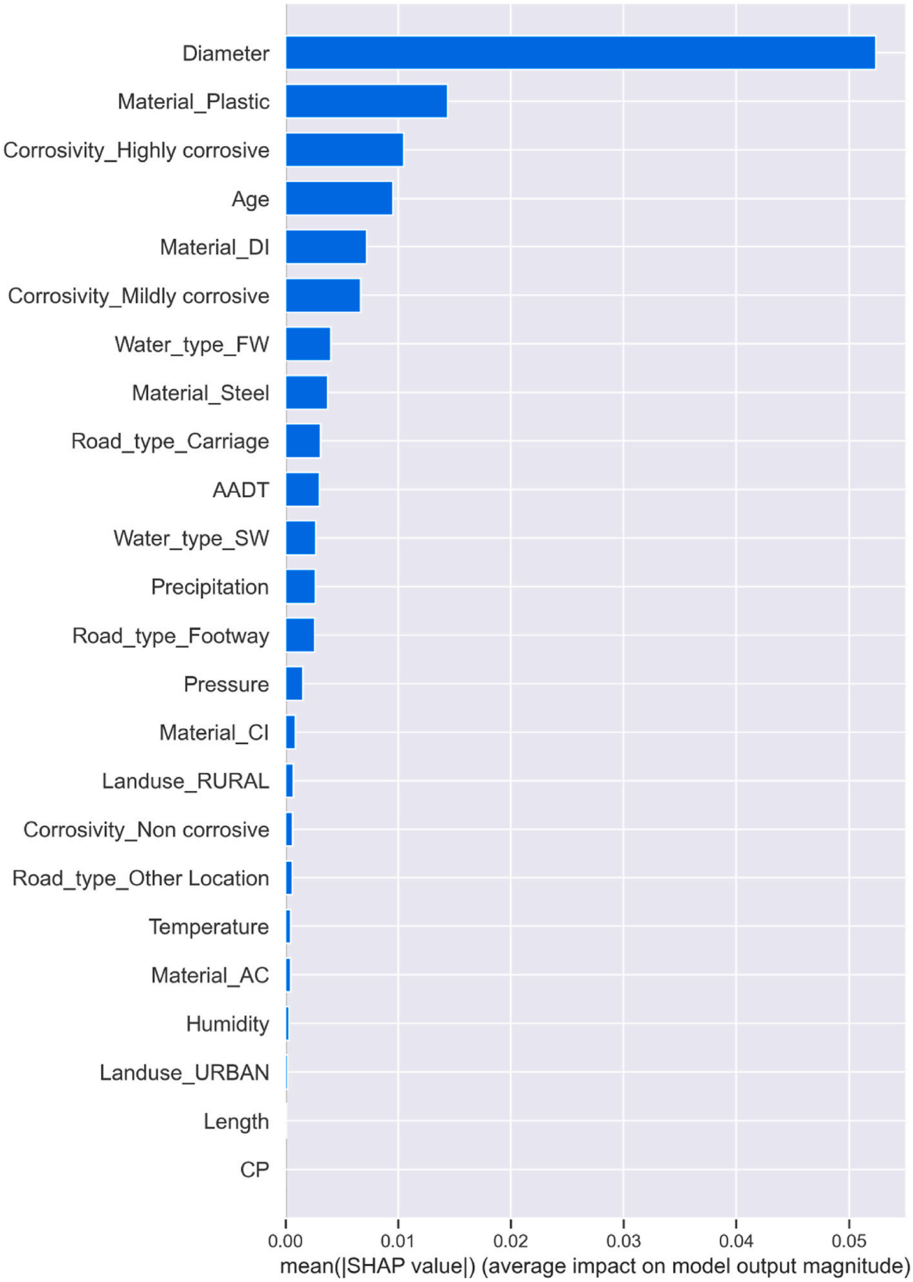


Fig. 7. SHAP feature importance for the probability of burst model.

strength and increased failure probability (Barton et al., 2019).

For age-related failures, the causal mechanism involves cumulative material degradation through various processes: mechanical fatigue from pressure cycling, chemical deterioration from water exposure, and environmental stress cracking. These processes follow well-established materials science principles and explain why older pipes consistently show higher failure probabilities (Farh et al., 2023). The high importance of 'Corrosivity_Highly corrosive' in burst prediction can be explained through electrochemical processes. In highly corrosive soils, increased ionic content accelerates electrochemical reactions at the pipe surface, leading to material loss through anodic dissolution. This process creates localised weak points that are more susceptible to bursts under pressure (Zhou et al., 2022).

Furthermore, the figures reveal that the presence of saltwater (denoted by 'Water_type_SW') and higher pressure levels are positively correlated with the likelihood of both leaks and bursts. This positive correlation implies that the model recognises these conditions as risk factors, with saltwater potentially accelerating electrochemical corrosion through pitting and crevice formation, while increased pressure directly increases mechanical stress on pipe walls and joints (Barton et al., 2019). The model's sensitivity to these variables underscores their importance in the predictive framework and potentially guides targeted maintenance efforts where these factors are prevalent.

6. Deployment of the POL and POB apps

After determining the optimised CNN as the most effective model based on the evaluations in Section 5.3, the next step was to deploy this model for predicting both POL and POB. The CNN model was configured to use standardised data for leak predictions and normalised data for burst predictions. To ensure a user-friendly and accessible deployment, the Streamlit framework was employed (Mhadbi, 2021). Streamlit allows for the development of interactive web applications with ease, providing a seamless platform for users to input data and receive

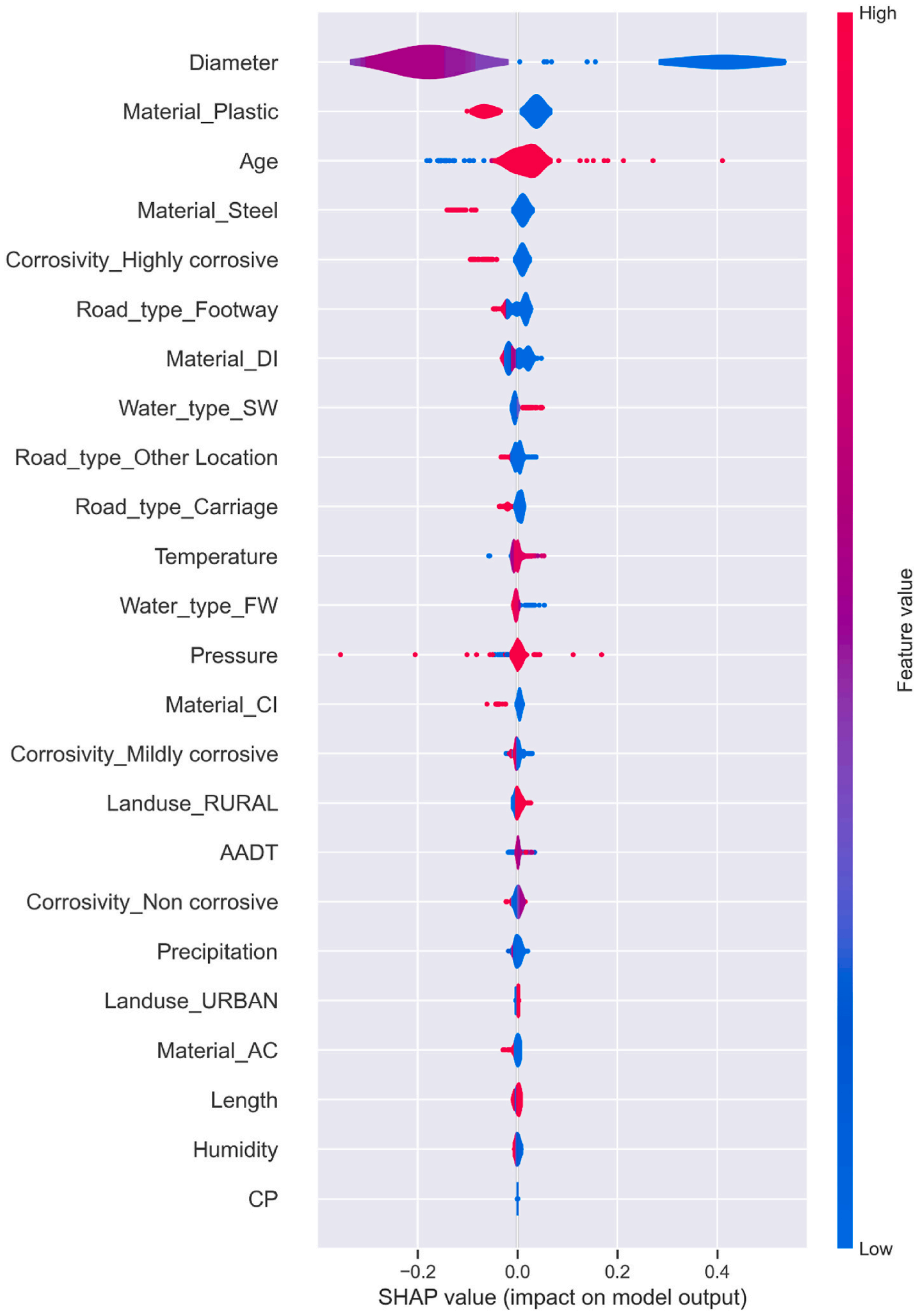


Fig. 8. Distribution of SHAP values for each feature for the probability of leak model.

predictions. Two separate web applications were developed - one for POL predictions and another for POB predictions, each maintaining similar user interface designs but tailored to their specific prediction tasks. For brevity, only screenshots of the POL prediction application are shown in Fig. 10.

The web application features two main prediction modes - single and batch prediction - accessible through clearly labelled tabs at the top of the interface. Fig. 10a presents a list of the 14 features required for prediction, meticulously organised with their respective units and possible values. The features include physical pipe characteristics (such as length in meters, diameter in millimeters), environmental factors (such as temperature in °C, relative humidity), operational parameters (such as pressure in bars), and categorical variables (material types including AC, CI, DI, GI, etc., corrosivity levels, road type, water type, and land use classification). This detailed feature list ensures that users understand the exact requirements and units for each input parameter.

Fig. 10b showcases the single prediction mode output, where users can input individual pipe parameters and receive a probability score for potential leakage. This interface is designed to predict specific pipes of interest and provide immediate feedback on their likelihood of failure. The probability output is presented in a clear, easily interpretable format.

Fig. 10c demonstrates the batch prediction functionality, a feature for analysing multiple pipes simultaneously. This interface comprises three main components: (1) a data source selection option allowing users to either utilise a pre-loaded test dataset or upload their own file, (2) a data preview table displaying the input parameters for multiple pipes in a structured format, and (3) a results section showing the predicted probabilities. The interface includes convenient functionality through the “Predict for Test Data” and “Download Results as CSV” buttons, enabling users to process multiple predictions and export results for further use. It is important to note that the actual prediction

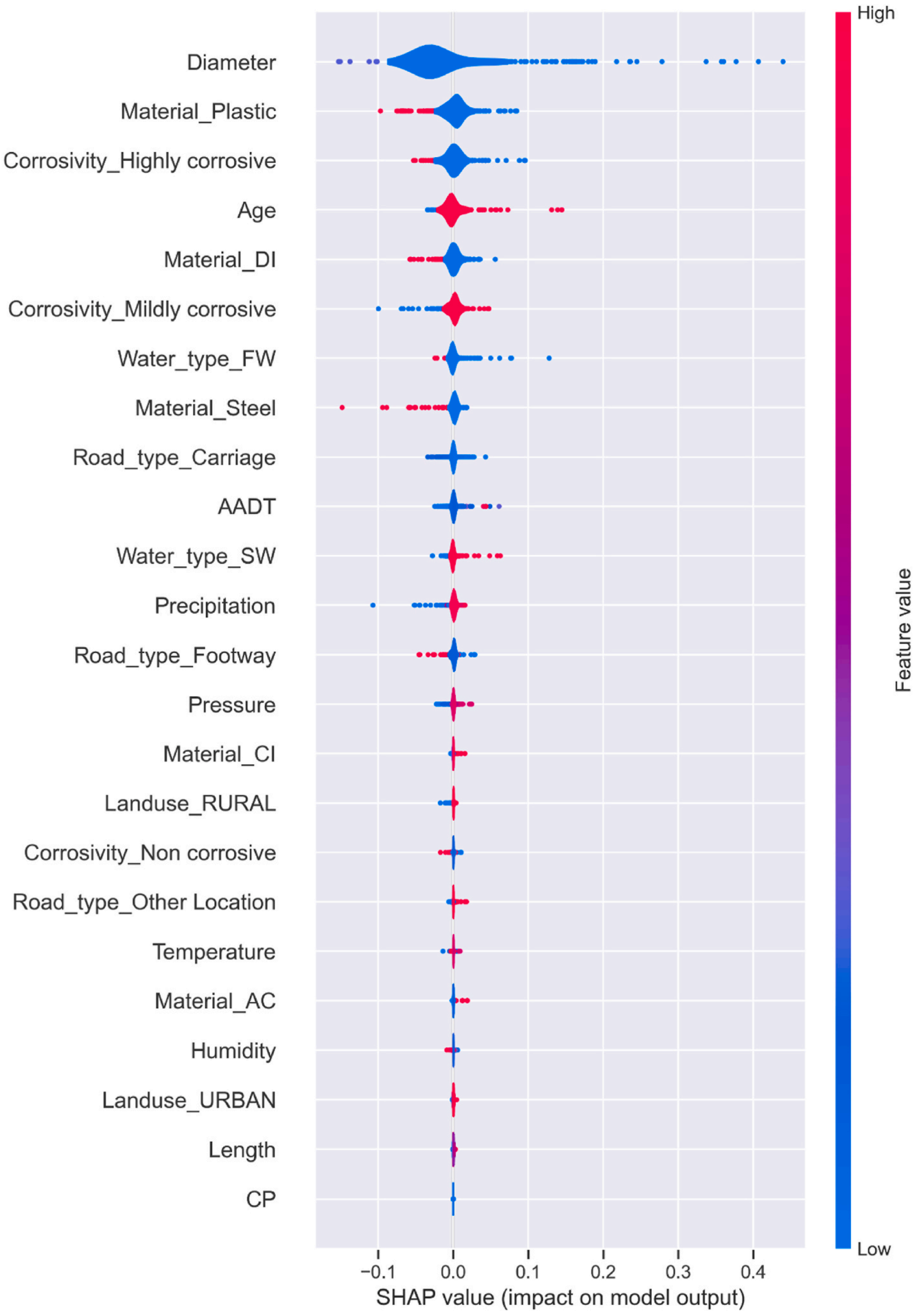


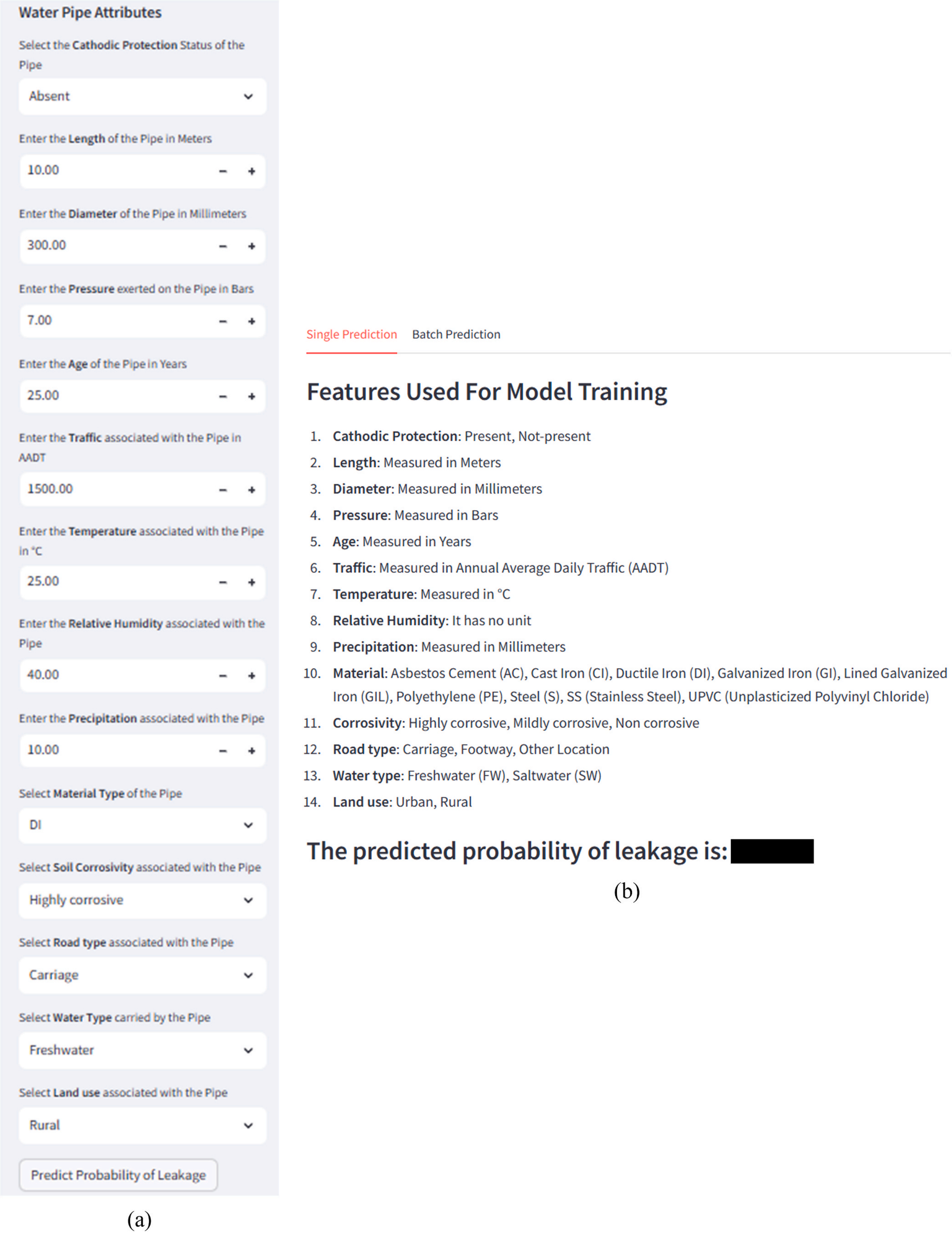
Fig. 9. Distribution of SHAP values for each feature for the probability of burst model.

results in Fig. 10a and b are blurred due to the confidential nature of the network data. This measure ensures the protection of sensitive information while still allowing the functionality of the models and web application to be demonstrated through the GUI. This approach maintains data security while effectively illustrating the application's capabilities and user interface design.

7. Conclusion

The degradation of water distribution networks (WDNs) carries significant environmental, economic, and societal costs. A critical strategy to mitigate these issues is the prediction of the probability of leaks and bursts — areas that remain relatively unexplored in existing research. This study has leveraged deep learning (DL) architectures, specifically deep neural networks (DNN), convolutional neural networks (CNN), and TabNet, to develop predictive models for the likelihood of leaks and bursts in water pipes. The study enhanced these base DL models by optimising their hyperparameters through Bayesian Optimisation, interpreting the optimal models with SHapley Additive exPlanations (SHAP), and deploying them as web applications.

Data compiled from the Hong Kong Water Supply Department, the Hong Kong Observatory, and the Hong Kong Transportation Department were categorised into variables associated with pipe characteristics, environmental conditions, and operational factors. Data processing involved removing outliers and imputing missing values through established statistical methods, followed by data normalisation and standardisation. It was observed that transforming the data remarkably augmented the models' predictive power. For example, in the context of leak prediction, the precision of the TabNet model increased from 0.512 to 0.598 with normalisation and further to 0.727 following standardisation. Additionally, the refinement of the models through optimisation led to enhanced performance compared to the base-DL models.



Water Pipe Attributes

Select the **Cathodic Protection** Status of the Pipe

Absent

Enter the **Length** of the Pipe in Meters

10.00

Enter the **Diameter** of the Pipe in Millimeters

300.00

Enter the **Pressure** exerted on the Pipe in Bars

7.00

Enter the **Age** of the Pipe in Years

25.00

Enter the **Traffic** associated with the Pipe in AADT

1500.00

Enter the **Temperature** associated with the Pipe in °C

25.00

Enter the **Relative Humidity** associated with the Pipe

40.00

Enter the **Precipitation** associated with the Pipe

10.00

Select **Material Type** of the Pipe

DI

Select **Soil Corrosivity** associated with the Pipe

Highly corrosive

Select **Road type** associated with the Pipe

Carriage

Select **Water Type** carried by the Pipe

Freshwater

Select **Land use** associated with the Pipe

Rural

(a)

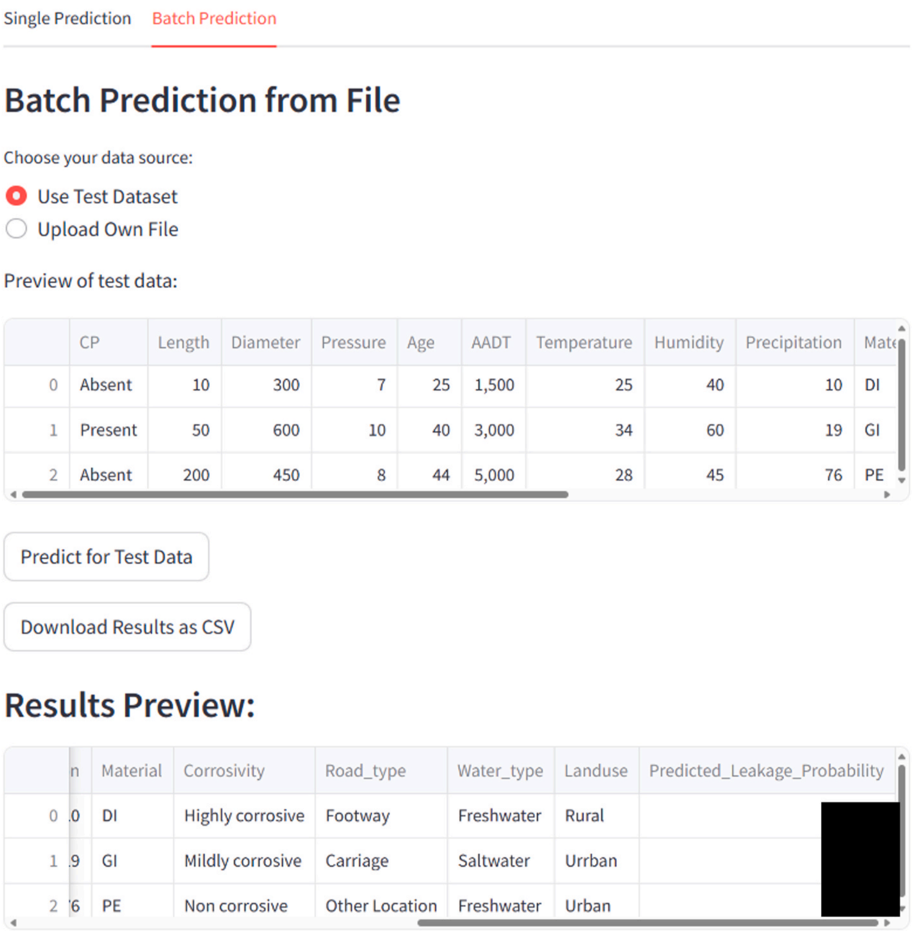
Features Used For Model Training

1. **Cathodic Protection**: Present, Not-present
2. **Length**: Measured in Meters
3. **Diameter**: Measured in Millimeters
4. **Pressure**: Measured in Bars
5. **Age**: Measured in Years
6. **Traffic**: Measured in Annual Average Daily Traffic (AADT)
7. **Temperature**: Measured in °C
8. **Relative Humidity**: It has no unit
9. **Precipitation**: Measured in Millimeters
10. **Material**: Asbestos Cement (AC), Cast Iron (CI), Ductile Iron (DI), Galvanized Iron (GI), Lined Galvanized Iron (GIL), Polyethylene (PE), Steel (S), SS (Stainless Steel), UPVC (Unplasticized Polyvinyl Chloride)
11. **Corrosivity**: Highly corrosive, Mildly corrosive, Non corrosive
12. **Road type**: Carriage, Footway, Other Location
13. **Water type**: Freshwater (FW), Saltwater (SW)
14. **Land use**: Urban, Rural

The predicted probability of leakage is: [redacted]

(b)

Fig. 10. GUI of the POL application: (a) Model inputs for a single pipe prediction; (b) Model output for a single pipe prediction; (c) Model inputs and output for batch predictions.



Batch Prediction from File

Choose your data source:

● Use Test Dataset
○ Upload Own File

Preview of test data:

	CP	Length	Diameter	Pressure	Age	AADT	Temperature	Humidity	Precipitation	Mate
0	Absent	10	300	7	25	1,500	25	40	10	DI
1	Present	50	600	10	40	3,000	34	60	19	GI
2	Absent	200	450	8	44	5,000	28	45	76	PE

Results Preview:

	n	Material	Corrosivity	Road_type	Water_type	Landuse	Predicted_Leakage_Probability
0	.0	DI	Highly corrosive	Footway	Freshwater	Rural	[redacted]
1	.9	GI	Mildly corrosive	Carriage	Saltwater	Urrban	[redacted]
2	'6	PE	Non corrosive	Other Location	Freshwater	Urban	[redacted]

(c)

Fig. 10. (*continued*).

This is exemplified by the increase in recall for the DNN model in burst prediction, which improved from 0.914 to 0.947 with normalisation and further to 0.950 after standardisation. The CNN was identified via the Copeland algorithm as the most proficient model for forecasting the probability of leaks and bursts. The SHAP analysis highlighted the predominance of 'diameter' and 'material composition, particularly plastic, in influencing model predictions.

The insights gained from this investigation are invaluable for proactive management of WDNs. The predictive models developed can help utility companies mitigate pipe failures and bolster the reliability of their supply infrastructure. Specifically, the probability scores from the algorithms can be used to prioritise pipes for preventative repair and replacement interventions. By combining the predicted failure likelihoods with information on pipe age, past repairs, and potential failure impacts, utility managers can make risk-informed decisions on capital investments. Focusing pipe rehabilitation efforts on high-risk pipes based on their modelled failure probabilities and failure consequences can optimise the effectiveness of infrastructure maintenance budgets.

The study acknowledges certain limitations, such as the models' performance being dependent on the quality and quantity of available data. Furthermore, while the models performed well in this study's context, their effectiveness might vary in different geographical locations or under various operational conditions. The study also faced challenges with the class imbalance in the dataset, which was addressed using SMOTE; however, future research could investigate other balancing techniques, such as class weight adjustment or random undersampling, to potentially improve model performance. Future investigations should focus on integrating more diverse datasets, including real-time monitoring data and pipe maintenance records, to further enhance model accuracy. Additionally, exploring the use of hybrid models that combine different DL architectures could offer new insights. Further research into the interpretability of these models is also crucial for practical applications, ensuring that utility managers can understand and trust the model predictions and use them to optimise infrastructure maintenance planning.

CRediT authorship contribution statement

Ridwan Taiwo: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Tarek Zayed:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization. **Beenish Bakhtawar:** Writing – review & editing, Data curation. **Bryan T. Adey:** Writing – review & editing, Methodology.

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.

Acknowledgement

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2025.124738.

Data availability

The authors do not have permission to share data.

References

Akinosho, T.D., Oyedele, L.O., Bilal, M., Ajayi, A.O., Delgado, M.D., Akinade, O.O., Ahmed, A.A., 2020. Deep learning in the construction industry: a review of present status and future innovations. J. Build. Eng. 32, 101827. https://doi.org/10.1016/j.jobe.2020.101827.

al-Ani, O., Das, S., Wu, H., 2023. Imitation learning with deep attentive tabular neural networks for environmental prediction and control in smart home. Energies 16 (13), 5091. https://doi.org/10.3390/en16135091.

Al-barqawi, H., Zayed, T., 2018. Infrastructure Management : Integrated AHP / ANN Model to Evaluate Municipal Water Mains ' Performance Infrastructure Management : Integrated AHP / ANN Model to Evaluate Municipal Water Mains ' Performance 342 (December 2008). https://doi.org/10.1061/(ASCE)1076-0342(2008)14.

Almheiri, Z., Meguid, M., Zayed, T., 2021. Failure modeling of water distribution pipelines using meta-learning algorithms. Water Res. 205 (September), 117680. https://doi.org/10.1016/j.watres.2021.117680.

Amiri-Ardakani, Y., Najafzadeh, M., 2021. Pipe break rate assessment while considering physical and operational factors: a methodology based on global positioning system and data-driven techniques. Water Resour. Manag., 0123456789 https://doi.org/10.1007/s11269-021-02911-6.

Arık, S., Pfister, T., 2021. TabNet: attentive interpretable tabular learning. In: 35th AAAI Conference on Artificial Intelligence, 8A. AAAI, pp. 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826, 2021.

Barton, N.A., Farewell, T.S., Hallett, S.H., Acland, T.F., 2019. Improving pipe failure predictions: factors effecting pipe failure in drinking water networks. Water Res. 164. https://doi.org/10.1016/j.watres.2019.114926.

Bello, I.T., Taiwo, R., Esan, O.C., Adegoke, A.H., Ijaola, A., Li, Z., Zhao, S., Chen, W., Shao, Z., Ni, M., 2024. AI-Enabled materials discovery for advanced ceramic electrochemical cells. Energy and AI 15, 100317. https://doi.org/10.1016/j.egyai.2023.100317.

Centers for Disease Control and Prevention, 2023. Waterborne Disease in the United States. Centers for Disease Control and Prevention. https://www.cdc.gov/healthywater/surveillance/burden/index.html.

Chen, M., Xu, Z., 2024. A deep learning classification framework for research methods of marine protected area management. J. Environ. Manag. 368 (April), 122228. https://doi.org/10.1016/j.jenvman.2024.122228.

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

Dawood, T., Elwakil, E., Novoa, H.M., Delgado, J.F.G., 2021. Ensemble intelligent systems for predicting water network condition index. Sustain. Cities Soc. 73 (January). https://doi.org/10.1016/j.scs.2021.103104.

Dillon, E., LaRiviere, J., Lundberg, S., Roth, J., Syrgkanis, V., 2018. Be careful when interpreting predictive models in search of causal insights. SHAP Documentation. Retrieved on 15/01/2025 from https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Becarefulwheninterpretingpredictivemodelsinsearchofcausalinsights.html.

Fahmy, M., Moselhi, O., 2009. Forecasting the remaining useful life of cast iron water mains. J. Perform. Constr. Facil. 23 (4), 269–275.

Fan, X., Wang, X., Zhang, X., Xiong, P.E.F.A., Yu, B., 2022. Machine learning based water pipe failure prediction : the effects of engineering , geology , climate and socio-economic factors. Reliab. Eng. Syst. Saf. 219 (November 2021), 108185. https://doi.org/10.1016/j.ress.2021.108185.

Farh, H.M.H., Ben Seghier, M.E.A., Taiwo, R., Zayed, T., 2023. Analysis and ranking of corrosion causes for water pipelines: a critical review. npj Clean Water, 6(65). https://doi.org/10.1038/s41545-023-00275-5.

Folkman, S., 2018. Water main break rates in the USA and Canada: A comprehensive study (Issue March). Retrieved on 20/04/2024 from Utah State University DigitalCommons via https://www.uni-bell.org/portals/0/ResourceFile/water_main_break_rates_in_the_usa_and_canada_a_comprehensive_study_march_2018.pdf.

Furxhi, I., Murphy, F., Mullins, M., Poland, C.A., 2019. Machine learning prediction of nanoparticle in vitro toxicity: a comparative study of classifiers and ensemble-classifiers using the Copeland Index. Toxicol. Lett. 312 (May), 157–166. https://doi.org/10.1016/j.toxlet.2019.05.016.

Geem, Z.W., Tseng, C.L., Kim, J., Bae, C., 2007. Trenchless water pipe condition assessment using artificial neural network. Pipelines 2007: Advances and Experiences with Trenchless Pipeline Projects - Proceedings of the ASCE International Conference on Pipeline Engineering and Construction 26. https://doi.org/10.1061/40934(252)26.

Giraldo-González, M.M., Rodríguez, J.P., 2020. Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks. Water (Switzerland) 12 (4). https://doi.org/10.3390/W12041153.

Harvey, R., Mcbean, E.A., Gharabaghi, B., 2014. Predicting the timing of water main failure using artificial neural networks. J. Water Resour. Plann. Manag. 140 (4), 425–434. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000354.

Ismaeel, M., Zayed, T., 2018. Integrated performance assessment model for water networks. J. Infrastruct. Syst. 24 (2), 04018005. https://doi.org/10.1061/(asce)is.1943-555x.0000419.

Jiang, R., Rathnayaka, S., Shannon, B., Zhao, X.L., Ji, J., Kodikara, J., 2019. Analysis of failure initiation in corroded cast iron pipes under cyclic loading due to formation of through-wall cracks. Eng. Fail. Anal. 103 (September 2018), 238–248. https://doi.org/10.1016/j.engfailanal.2019.04.031.

Jun, H.J., Park, J.K., Bae, C.H., 2017. Deep leaning neural networks for determining replacement timing of steel water transmission pipes. Proceedings - 2017 International Conference on Control, Artificial Intelligence, Robotics and Optimization, ICCAIRO 2017 2018-Janua, 219–225. https://doi.org/10.1109/ICCAIRO.2017.49.

Kabir, G., Tesfamariam, S., Francisque, A., Sadiq, R., 2015. Evaluating risk of water mains failure using a Bayesian belief network model. Eur. J. Oper. Res. 240 (1), 220–234. https://doi.org/10.1016/j.ejor.2014.06.033.

Karamouz, M., Yousefi, M., Zahmatkesh, Z., Nazif, S., 2012. Development of an algorithm for vulnerability zoning of water distribution network. World Environmental and Water Resources Congress 2012: Crossing Boundaries 3011–3020.

Kerwin, S., Adey, B.T., 2020. Optimal intervention planning: a bottom-up approach to renewing aging water infrastructure. J. Water Resour. Plann. Manag. 146 (7). https://doi.org/10.1061/(asce)wr.1943-5452.0001217.

Kerwin, S., Adey, B.T., 2021. Exploiting digitalisation to plan interventions on large water distribution networks. Infrastructure Asset Management 9 (4), 207–222. https://doi.org/10.1680/jinam.20.00017.

Kerwin, S., Garcia de Soto, B., Adey, B., Sampatakaki, K., Heller, H., 2020. Combining recorded failures and expert opinion in the development of ANN pipe failure prediction models. Sustainable and Resilient Infrastructure 8 (1), 1–23. https://doi.org/10.1080/23789689.2020.1787033.

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

Kutyłowska, M., 2015. Neural network approach for failure rate prediction. Eng. Fail. Anal. 47, 41–48. https://doi.org/10.1016/j.engfailanal.2014.10.007.

Li, Z., Liu, H., Zhang, C., Fu, G., 2024. Gated graph neural networks for identifying contamination sources in water distribution systems. J. Environ. Manag. 351 (November 2023), 119806. https://doi.org/10.1016/j.jenvman.2023.119806.

Liu, S., Gunawan, C., Barraud, N., Rice, S.A., Harry, E.J., Amal, R., 2016. Understanding, monitoring, and controlling biofilm growth in drinking water distribution systems. Environ. Sci. Technol. 50 (17), 8954–8976. https://doi.org/10.1021/acs.est.6b00835.

Mazumder, R.K., Salman, A.M., Li, Y., 2021. Failure risk analysis of pipelines using data-driven machine learning algorithms. Struct. Saf. 89 (July 2020), 102047. https://doi.org/10.1016/j.strusafe.2020.102047.

Mhadbi, N., 2021. Python Tutorial: Streamlit. Datacamp. https://www.datacamp.com/tutorial/streamlit.

Mian, H.R., Hu, G., Hewage, K., Rodriguez, M.J., Sadiq, R., 2023. Drinking water management strategies for distribution networks: an integrated performance assessment framework. J. Environ. Manag. 325 (PB), 116537. https://doi.org/10.1016/j.jenvman.2022.116537.

Nguyen, H.V., Byeon, H., 2023. Predicting depression during the COVID-19 pandemic using interpretable TabNet: a case study in South Korea. Mathematics 11 (14). https://doi.org/10.3390/math11143145.

Pękala, A., Pietrucha-Urbanik, K., 2018. The influence of the soil environment on the corrosivity of failure infrastructure - case study of the exemplary water network. Arch. Civ. Eng. 64 (1), 133–144. https://doi.org/10.2478/ace-2018-0009.

Raziani, S., Azimbagirad, M., 2022. Deep CNN hyperparameter optimisation algorithms for sensor-based human activity recognition. Neuroscience Informatics 2 (3), 100078. https://doi.org/10.1016/j.neuri.2022.100078.

Rezaei, H., Ryan, B., Stoianov, I., 2015. Pipe failure analysis and impact of dynamic hydraulic conditions in water supply networks. Procedia Eng. 119 (1), 253–262. https://doi.org/10.1016/j.proeng.2015.08.883.

Rifaai, T.M., Abokifa, A.A., Sela, L., 2022. Integrated approach for pipe failure prediction and condition scoring in water infrastructure systems. Reliab. Eng. Syst. Saf. 220 (December 2021), 108271. https://doi.org/10.1016/j.ress.2021.108271.

Robles-velasco, A., Cortés, P., Muñuzuri, J., Onieva, L., Organización, D. De, Empresas, G. De, Etsi, I.I., Sevilla, U. De, Descubrimientos, C.C.D.L., 2020. Prediction of pipe failures in water supply networks using logistic regression and support vector classification. Reliab. Eng. Syst. Saf. 196 (March 2019), 106754. https://doi.org/10.1016/j.ress.2019.106754.

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

Snider, B., McBean, E.A., 2018. Improving time-To-failure predictions for water distribution systems using gradient boosting algorithm. In: 1st International WDSA/CCWI 2018 Joint Conference, July.

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

Taiwo, R., Ben Seghier, M.E.A., Zayed, T., 2023c. Predicting wall thickness loss in water pipes using machine learning techniques. 2nd Conference of the European Association on Quality Control of Bridges and Structures - EUROSTRUCT2023 6 (5), 1087–1092. https://doi.org/10.1002/cepa.2075.

Taiwo, R., Ben Seghier, M.E.A., Zayed, T., 2023b. Towards sustainable water infrastructure : the state-of-the-art for modeling the failure probability of water pipes. Water Resour. Res. 59 (4), e2022WR033256. https://doi.org/10.1029/2022WR033256.

Taiwo, R., Shaban, I.A., Zayed, T., 2023a. Development of sustainable water infrastructure: a proper understanding of water pipe failure. J. Clean. Prod. 398, 136653. https://doi.org/10.1016/j.jclepro.2023.136653.

Taiwo, R., Yussif, A.M., Ben Seghier, M.E.A., Zayed, T., 2024b. Explainable ensemble models for predicting wall thickness loss of water pipes. Ain Shams Eng. J., 102630 https://doi.org/10.1016/j.asej.2024.102630. January.

Taiwo, R., Zayed, T., Ben Seghier, M.E.A., 2024a. Integrated intelligent models for predicting water pipe failure probability. Alex. Eng. J. 86, 243–257. https://doi.org/10.1016/j.aej.2023.11.047.

Tariq, S., Bakhtawar, B., Zayed, T., 2022. Data-driven application of MEMS-based accelerometers for leak detection in water distribution networks. Sci. Total Environ. 809, 151110. https://doi.org/10.1016/j.scitotenv.2021.151110.

Tariq, S., Hu, Z., Zayed, T., 2021. Micro-electromechanical systems-based technologies for leak detection and localisation in water supply networks: a bibliometric and systematic review. J. Clean. Prod. 289, 125751. https://doi.org/10.1016/j.jclepro.2020.125751.

Tavakoli, R., Sharifara, A., Najafi, M., 2020. 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, vol. 2001. ASCE, pp. III–IV.

Tsai, Y.L., Chang, H.C., Lin, S.N., Chiou, A.H., Lee, T.L., 2022. Using convolutional neural networks in the development of a water pipe leakage and location identification system. Appl. Sci. 12 (16). https://doi.org/10.3390/app12168034.

Uddin, M.G., Nash, S., Mahammad Diganta, M.T., Rahman, A., Olbert, A.I., 2022. Robust machine learning algorithms for predicting coastal water quality index. J. Environ. Manag. 321 (June), 115923. https://doi.org/10.1016/j.jenvman.2022.115923.

Wasim, M., Shoaib, S., Mubarak, N.M., Inamuddin, Asiri, A.M., 2018. Factors influencing corrosion of metal pipes in soils. Environ. Chem. Lett. 16 (3), 861–879. https://doi.org/10.1007/s10311-018-0731-x.

Water Supplies Department HKSAR, 2021. WSD annual report. https://www.wsd.gov.hk/filemanager/common/annual_report/2019_20/en/index.html.

Weeraddana, D., MallawaArachchi, S., Warnakula, T., Li, Z., Wang, Y., 2021. Long-term pipeline failure prediction using nonparametric survival analysis. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12460 LNAI 139–156. https://doi.org/10.1007/978-3-030-67667-4_9.

Wilson, D., Filion, Y., Moore, I., 2017. State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. Urban Water J. 14 (2), 173–184. https://doi.org/10.1080/1573062X.2015.1080848.

Winkler, D., Haltmeier, M., Kleidorfer, M., Rauch, W., Tscheikner-Gratl, F., 2018. Pipe failure modelling for water distribution networks using boosted decision trees. Structure and Infrastructure Engineering 14 (10), 1402–1411. https://doi.org/10.1080/15732479.2018.1443145.

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

Yang, L., Shami, A., 2020. On hyperparameter optimisation of machine learning algorithms: theory and practice. Neurocomputing 415, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061.

Yeung, H.C., Ridwan, T., Tariq, S., Zayed, T., 2020. BEAM Plus implementation in Hong Kong: assessment of challenges and policies. International Journal of Construction Management. https://doi.org/10.1080/15623599.2020.1827692.

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

Zheng, Y., Wei, J., Zhang, W., Zhang, Y., Zhang, T., Zhou, Y., 2024. An ensemble model for accurate prediction of key water quality parameters in river based on deep learning methods. J. Environ. Manag. 366 (July), 121932. https://doi.org/10.1016/j.jenvman.2024.121932.

Zhou, H., Li, P., Wu, L., 2022. Research on a model-based burst pressure prediction method for pipelines with corrosion defects. Journal of Mechanics 38, 315–322.