

Article

A Stacking-Based Ensemble Learning Framework for Water Distribution Network Condition Prediction and Interpretability

Qingfu Li and Ao Chen *

School of Water Conservancy and Transportation, Zhengzhou University, Zhengzhou 450001, China;
lqflch@zzu.edu.cn

* Correspondence: aochen@gs.zzu.edu.cn

Abstract

Water distribution network (WDN) pipelines are essential infrastructure, and their failures cause significant economic and operational losses. Existing condition prediction models often struggle with high-dimensional non-linearities, severe class imbalance, and a lack of decision-centric interpretability. To address these challenges, this study presents a Stacking-based ensemble learning framework that integrates Random Forest, XGBoost, and LightGBM base learners with a Logistic Regression meta-learner under a spatially constrained group sampling scheme. Model performance is evaluated alongside an independent CatBoost benchmark using macro-averaged metrics, an operational cost-loss function for high-risk assets, and Copeland ranking, supplemented by global and class-specific SHAP interpretability analyses. Evaluated on a pipeline dataset, the Stacking framework achieved superior overall performance and recorded the lowest operational cost loss by minimizing severe misclassifications of critical pipelines. SHAP analysis identified pipe age, material, physical dimensions, temperature, and soil moisture as the primary risk drivers. This integrated framework delivers a robust, transparent, and decision-centric decision support tool for proactive municipal pipeline maintenance and risk mitigation.

Keywords: water distribution networks; stacking ensemble learning; Copeland; SHAP; infrastructure

1. Introduction

The water distribution network (WDN) is a critical infrastructure essential for ensuring the reliable supply of potable water to urban and rural populations. Comprising pipelines, valves, pumps, and storage tanks, the network collaboratively delivers treated water from treatment plants to end users [1–4]. Among these components, pipelines constitute the vast majority of the physical assets and play a foundational role in water transmission. Consequently, pipeline failures can result in substantial economic losses, service interruptions, environmental degradation, and public safety concerns, underscoring the imperative for reliable condition assessment and predictive maintenance strategies [5].

Existing approaches for pipeline condition prediction can be broadly categorized into physics-based, statistical, and data-driven models [6]. Physics-based models aim to simulate the physical mechanisms leading to pipe failure by analyzing structural loads and material resistance [7]. While they offer strong interpretability by explicitly represent-

Received: 30 June 2026

Revised: 24 July 2026

Accepted: 27 July 2026

Published: 28 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

ing the relationships among environmental conditions, pipe characteristics, and failure mechanisms, these models typically demand extensive hydraulic, structural, and environmental data. This high data requirement renders their implementation costly and limits their applicability primarily to critical assets or large transmission mains [5,8].

Statistical models, in contrast, utilize historical failure records and pipeline condition attributes to estimate future deterioration trends, spanning deterministic methods (e.g., regression analysis) and probabilistic methods (e.g., Weibull-based hazard models) [9,10]. Previous studies have proposed several enhanced statistical approaches, including K-means clustering combined with Evolutionary Polynomial Regression [11], Bayesian Weibull Proportional Hazard Models [12], and life-cycle-based cumulative Weibull functions [13]. Although these methods effectively exploit historical failure patterns, they inherently struggle to capture complex nonlinear interactions among multiple influencing factors.

To improve inspection efficiency while reducing operational costs, researchers have increasingly explored data-driven approaches for pipeline condition assessment. Early machine learning studies demonstrated that Artificial Neural Networks (ANNs) could effectively capture the complex relationships between observable indicators and pipeline conditions, providing a promising alternative to traditional inspection methods [14]. Building upon these early efforts, the rapid advancement of machine learning has led to the widespread adoption of more sophisticated data-driven models for pipeline condition prediction owing to their superior capability to capture high-dimensional nonlinear relationships. For instance, Elshaboury and Marzouk [15] used various machine learning algorithms, including feedforward and general regression neural networks (GRNN), multiple linear regression, and support vector regression, to predict the structural condition of water pipes. Fan et al. [16] combined Long Short-Term Memory (LSTM) networks with Monte Carlo simulation for uncertainty quantification of failure rates. Jafari et al. [17] applied hybrid neuro-fuzzy models, demonstrating improved accuracy over conventional methods. Zali et al. [18] developed semi-supervised clustering methods to address class imbalance, while Liu et al. [19] and Robles-Velasco et al. [20] successfully applied deep learning and multi-label classification techniques, respectively. Other studies have employed logistic regression, support vector machines, and Random Forest (RF), occasionally incorporating feature importance analyses [21–23]. Despite their predictive advantages, standard data-driven models face persistent challenges, most notably the “black-box” nature, which limits their interpretability, high sensitivity to hyperparameters, and a tendency to overfit when applied to sparse or imbalanced datasets.

Beyond pipeline deterioration prediction, machine learning has also been increasingly applied to a wide range of water infrastructure problems. Recent studies have employed machine learning techniques to estimate viscoelastic creep-function parameters from transient hydraulic signals, optimize viscoelastic pipeline parameters using transient pressure responses, and support the optimal design of WDNs by integrating hydraulic characteristics with graph-theoretic features [24–27]. In addition, machine learning has been extensively adopted for resilience assessment, hydraulic simulation, sensor placement optimization, and leak localization in WDNs [28–30]. These studies demonstrate that machine learning has evolved beyond conventional failure prediction toward hydraulic parameter estimation, intelligent system analysis, and network-level optimization, reflecting its expanding role in modern water infrastructure management. Nevertheless, accurate and explainable pipeline condition assessment remains an important yet insufficiently explored research area.

To overcome the limitations of individual machine learning models, ensemble learning has attracted considerable attention. By combining multiple base learners, ensemble methods significantly improve predictive robustness, stability, and generalization capability [31,32]. Depending on the integration mechanism, these approaches are categorized

into bagging, boosting, and Stacking. Bagging methods, such as RF, enhance stability via bootstrap aggregation [33], while boosting methods (e.g., AdaBoost, XGBoost) sequentially correct the errors of preceding learners [34–36]. Stacking, originally proposed by Wolpert [37], employs a meta-learner to optimally combine predictions from diverse base learners. Although Stacking has shown strong potential in broader engineering applications, such as power outage and energy recovery prediction [38–40], its application to pipeline condition modeling—especially when paired with rigorous interpretability frameworks—warrants further investigation.

Despite significant progress in pipeline condition prediction, critical research gaps persist in the existing literature. First, while various machine learning techniques have been applied, the exploration of robust ensemble strategies remains limited. Most studies rely on single, isolated models that are inherently susceptible to data volatility or algorithm-specific biases. Consequently, current approaches often fail to fully exploit the complementary strengths of heterogeneous algorithms, hindering the development of highly stable and robust predictive frameworks essential for practical deployment. Second, the interpretability of data-driven models for pipeline deterioration remains insufficient. Although recent research has incorporated basic interpretability tools, systematic investigations into how intrinsic structural properties interact with extrinsic environmental dynamics are still lacking. This constrains the practical utility of these models, as water authorities require transparent, causal insights into risk drivers rather than mere statistical correlations to guide proactive engineering interventions.

To address these limitations, this study develops a holistic, integration-based framework tailored for robust and decision-centric pipeline condition prediction. The specific objectives of this research are threefold:

1. To construct a Stacking-based ensemble learning framework that integrates heterogeneous tree-based base models (RF, XGBoost, and LightGBM) with a meta-learner (Logistic Regression) under a spatially constrained group sampling strategy to enhance prediction robustness and mitigate class imbalance in WDNs.
2. To establish a multi-dimensional evaluation and decision-centric ranking mechanism incorporating macro-averaged metrics, class-specific operational cost-loss functions for high-risk assets (red class), and the Copeland ranking algorithm to systematically compare the integrated ensemble against individual benchmark models.
3. To unveil the underlying risk drivers and nonlinear interactions across distinct pipeline degradation states (red, amber, and green) by applying global and class-specific SHAP analyses, providing causal and transparent insights for proactive engineering maintenance.

The proposed framework is empirically validated using a comprehensive, real-world dataset comprising 97,696 pipeline assets. By balancing model stability with deep transparency, this study provides a reliable, data-driven solution that directly supports risk mitigation and maintenance decision-making in water distribution systems.

2. Research Methodology

The methodology used in this study is shown in Figure 1.

2.1. Model Algorithms

2.1.1. Random Forest (RF)

RF is an ensemble learning algorithm that constructs multiple decision trees and aggregates their predictions to improve classification performance. Each decision tree is trained using bootstrap sampling from the original dataset, while a random subset of fea-

tures is selected at each node split. This randomization increases the diversity among trees and reduces the risk of overfitting compared with a single decision tree [41].

Due to its robustness, ability to capture nonlinear relationships, and strong performance on high-dimensional data, RF has been widely used in various classification problems. For classification tasks, the final prediction is obtained through majority voting across all trees:

$$\hat{y} = \arg \max_c \sum_{t=1}^T I(h_t(x) = c) \quad (1)$$

where $I(\cdot)$ is the indicator function, $h_t(x)$ denotes the prediction of the t -th decision tree, T is the total number of trees in the forest, and c represents a candidate class label. The final prediction $\hat{y}$ is determined by selecting the class receiving the highest number of votes from all trees.

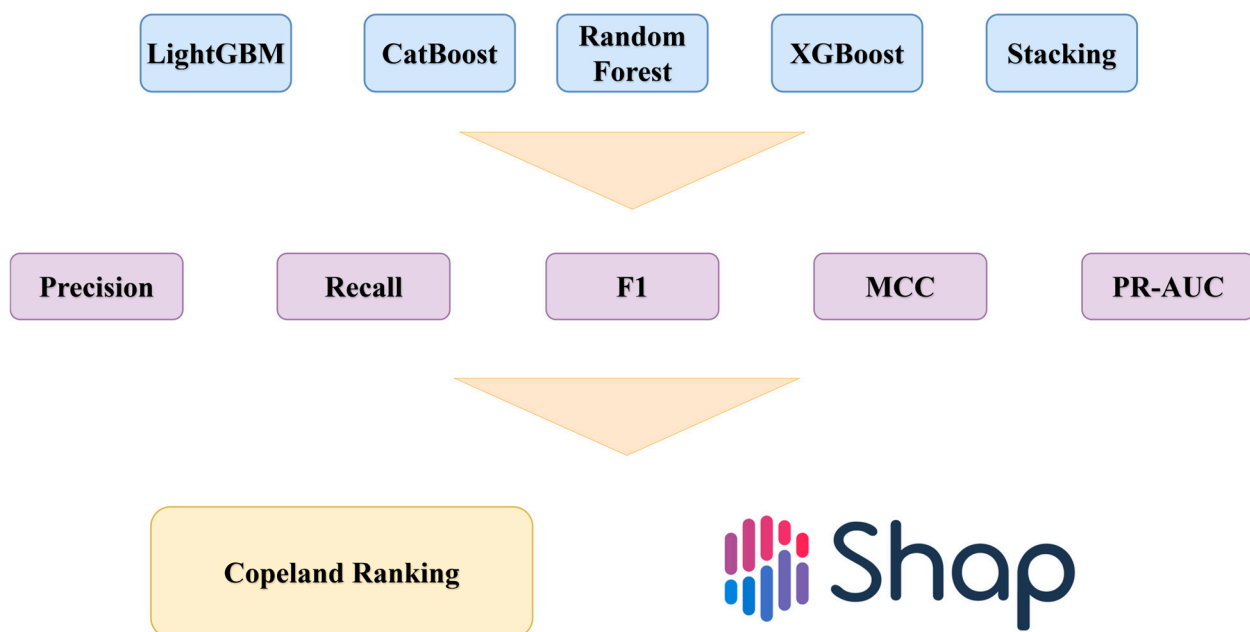


Figure 1. Study framework.

2.1.2. Gradient Boosting Decision Tree (GBDT)

Gradient boosting decision tree (GBDT) models construct an ensemble of decision trees in a sequential manner, where each newly added tree is trained to reduce the prediction errors of the existing ensemble. In this study, three representative GBDT-based algorithms, including XGBoost, LightGBM, and CatBoost, were employed. These algorithms share the same additive modeling principle, where the final prediction is obtained by combining the outputs of multiple decision trees:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (2)$$

where x_i denotes the input feature vector of the i -th sample, K represents the number of decision trees, and $f_k(\cdot)$ denotes the prediction function of the k -th tree. During training, each tree is sequentially optimized to minimize the overall objective function:

$$\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \Omega(f_k) \quad (3)$$

where $l(\cdot)$ represents the loss function and $\Omega(\cdot)$ denotes the regularization term used to control model complexity.

Although based on the same boosting principle, the three algorithms employ different optimization strategies. XGBoost improves model performance by introducing sec-

ond-order gradient information and explicit regularization. LightGBM enhances computational efficiency through histogram-based feature discretization and leaf-wise tree growth. CatBoost adopts ordered boosting and specialized categorical feature encoding strategies to reduce prediction bias and improve generalization capability.

2.1.3. Stacking

Stacking is a model-agnostic ensemble learning technique that combines multiple base models by training a meta-model to aggregate their predictions. In this framework, several base models are first trained on the training data. The predictions of these base models are then used as input features for the meta-model, which learns to optimally combine the base model outputs. Figure 2 is a schematic diagram of the Stacking process.

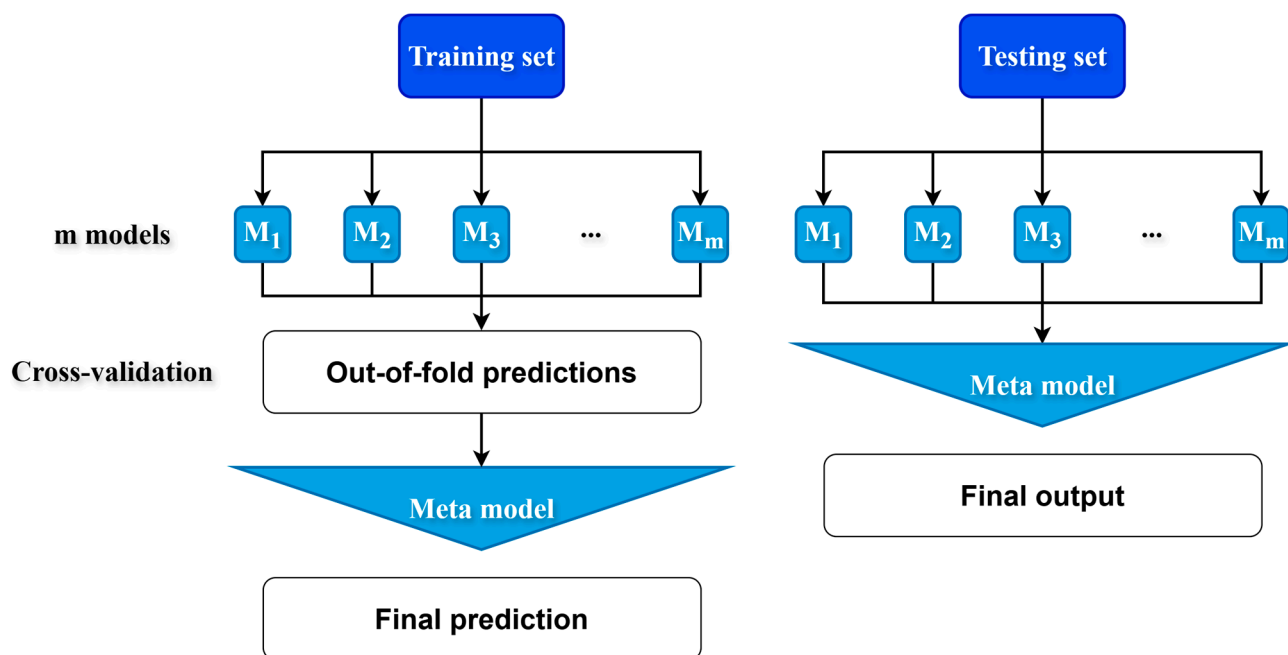


Figure 2. Stacking framework.

2.2. Performance Metrics

In this study, model performance is evaluated using macro-averaged precision, macro-averaged recall, macro-averaged F1-score, macro-averaged PR AUC, Matthews correlation coefficient (MCC), and class-specific precision, recall, F1-score, and PR AUC for the critical condition class.

Macro-averaged metrics are employed to assess the overall classification capability across all pipe condition categories by assigning equal importance to each class, thereby preventing the evaluation from being dominated by majority classes. The macro-averaged PR AUC further measures the model's discriminative ability under different decision thresholds, which is particularly informative for imbalanced classification tasks.

MCC is additionally adopted as a comprehensive evaluation metric because it considers all components of the confusion matrix and provides a balanced assessment of classification performance, even when class distributions are uneven.

Considering the practical objective of pipeline condition assessment, accurately identifying critical-condition pipes is of primary importance for maintenance prioritization and risk mitigation. Therefore, class-specific precision, recall, F1-score, and PR AUC are further reported for the critical class to evaluate the model's capability to detect high-risk pipelines. These indicators provide direct insights into whether the developed model can effectively support engineering decision-making and targeted maintenance strategies.

The definitions of precision, recall, F1-score, PR AUC, and MCC are presented below. Macro-averaged metrics are calculated by averaging the corresponding class-wise values, whereas class-specific metrics describe the prediction performance for the critical condition category.

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (4)$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \quad (5)$$

$$PR\ AUC = \int_0^1 Precision(Recall) dRecall \quad (6)$$

$$F1_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \quad (7)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

where TP_i , FP_i , and FN_i denote the numbers of true positives, false positives, and false negatives for class i , respectively; $Precision_i$, $Recall_i$, and $F1_i$ represent the precision, recall, and F1-score of class i , respectively.

2.3. Copeland Algorithm for Model Ranking

To obtain a comprehensive ranking of the evaluated models across multiple performance metrics, the Copeland method was adopted. The Copeland method is a pairwise comparison-based ranking approach widely used in decision-making and voting systems [42]. In this method, each model is compared with every other model based on its performance on the selected evaluation metrics.

For each pairwise comparison, a model receives a score of +1 if it outperforms the competing model, −1 if it performs worse, and 0 in the case of a tie. The Copeland score of a model is calculated as the difference between the total number of wins and losses across all pairwise comparisons. The final ranking is then determined according to the Copeland scores, where a higher score indicates better overall performance.

By aggregating results from multiple metrics, the Copeland method provides a robust and objective mechanism for identifying the most competitive models.

2.4. SHapley Additive exPlanations (SHAP)

To interpret the contribution of input features to the model predictions, SHAP was employed. SHAP is an explainable artificial intelligence method derived from Shapley values in cooperative game theory, which fairly allocates the contribution of each feature to the model output. The Shapley value represents the average marginal contribution of a feature across all possible feature subsets, ensuring desirable theoretical properties such as consistency and local accuracy [43,44].

The SHAP value for feature i is computed as shown in Equation (9):

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x) - f_S(x)] \quad (9)$$

where ϕ_i denotes the SHAP value of feature i , F represents the set of all input features, S is a subset of features excluding feature i , $|S|$ is the number of features in subset S , and $f_S(x)$ denotes the model output when only the features in subset S are considered. The term $f_{S \cup \{i\}}(x)$ represents the marginal contribution of feature i to the model prediction.

To evaluate the global importance of features across the dataset, the mean absolute SHAP value was used. This measure reflects the average magnitude of the contribution of a feature to the model predictions and is calculated as shown in Equation (10):

$$Importance_i = \frac{1}{N} \sum_{j=1}^N |\phi_{ij}| \quad (10)$$

where $Importance_i$ denotes the global importance of feature i , N is the total number of samples, and ϕ_{ij} represents the SHAP value of feature i for sample j .

3. Data Pre-Processing

The pipeline network dataset used in this study was obtained from the Zenodo research repository [45]. The initial database comprised 97,906 raw records. During the data preprocessing phase, rigorous quality control checks were first executed to ensure strict logical and physical consistency. Specifically, entries containing missing values or physically anomalous observations—such as negative pipeline age or negative structural length—were systematically excluded. Additionally, non-informative attributes, such as unit descriptors, were discarded to streamline the feature space. In total, 210 invalid records were removed during this filtration process, yielding a clean dataset of 97,696 valid samples.

To optimize feature representation and mitigate information redundancy, target-oriented feature engineering was performed. Pipeline service age was explicitly derived as the operational duration between the inspection year and the installation year, providing a structurally grounded indicator of cumulative asset deterioration compared with the original temporal variables. Regarding environmental variables, multiple statistical descriptors (including minimum, maximum, mean, and standard deviation) were initially evaluated. However, preliminary Pearson and Spearman correlation analyses revealed pervasive, high-degree collinearity among these descriptors. To circumvent potential multicollinearity that could destabilize base-learner training, only the long-term mean values were retained, thereby maintaining representative environmental drivers while preserving a parsimonious feature architecture.

Following preprocessing and variable selection, the finalized dataset comprised 97,696 valid pipeline records characterized by eight predictive variables and one target condition label. The selected features were symmetrically organized into four pipe-related attributes (length, diameter, service age, and material) and four external environmental factors (average soil moisture at depths of 40–100 cm and 100–200 cm, average air temperature, and annual average precipitation). Table 1 provides the summary statistics of the data, while Figure 3 displays the distribution of material types. Except for pipe material, which was treated as a categorical feature, all predictors were continuous numerical variables. To facilitate subsequent model development, the categorical material attribute was transformed using one-hot encoding, whereas all continuous variables were standardized via z-score normalization. To prevent information leakage, both the encoding and standardization parameters were learned exclusively from the training data and then consistently applied to the testing data. Inter-feature dependencies and multi-variable alignment were systematically validated via Pearson and Spearman correlation analyses, as illustrated in Figures 4 and 5, respectively.

Table 1. Summary statistics of the data.

Variable	Unit	Mean	Standard Deviation
Age	Years	32.4	24.0
Diameter	mm	127.2	136.9
Length	m	44.0	106.9
Material	AC, PVC, DI, ST, CI, CU, PE, CO, HDPE, OPVC, PVCB, MDPE, MDPE80, UPVC, MDPE100, SS, GS,	-	-

STCL, RCON, PE80, DICL, MPVC, PE100, PB, CICL, ABS, BUTYL, GI			
Mean soil moisture (40–100 cm)	mm	154.4	22.1
Mean soil moisture (100–200 cm)	mm	269.4	10.9
Mean air temperature	K	283.5	2.5
Mean annual rainfall	mm/year	336.3	94.3

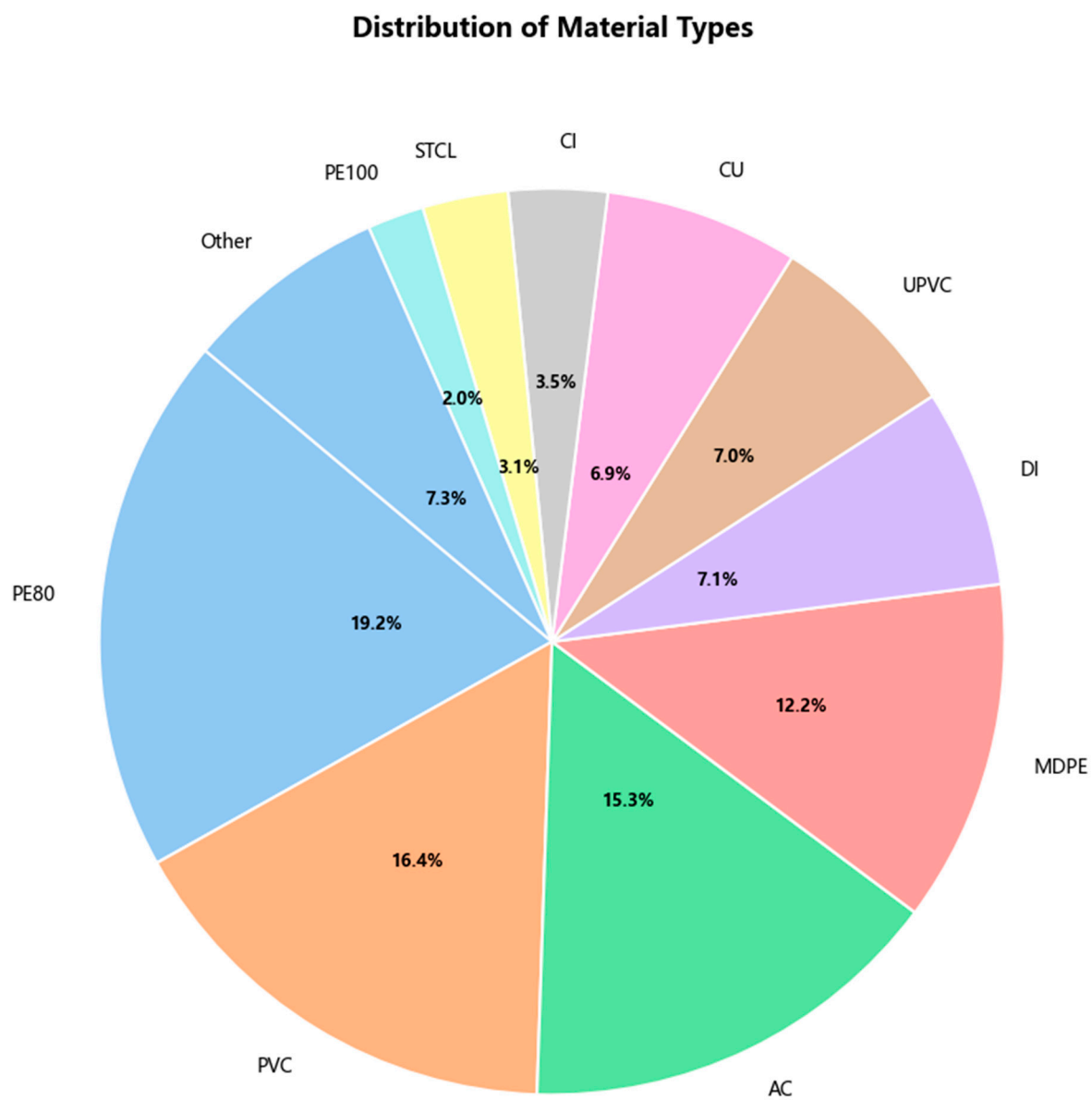


Figure 3. Distribution of material types.

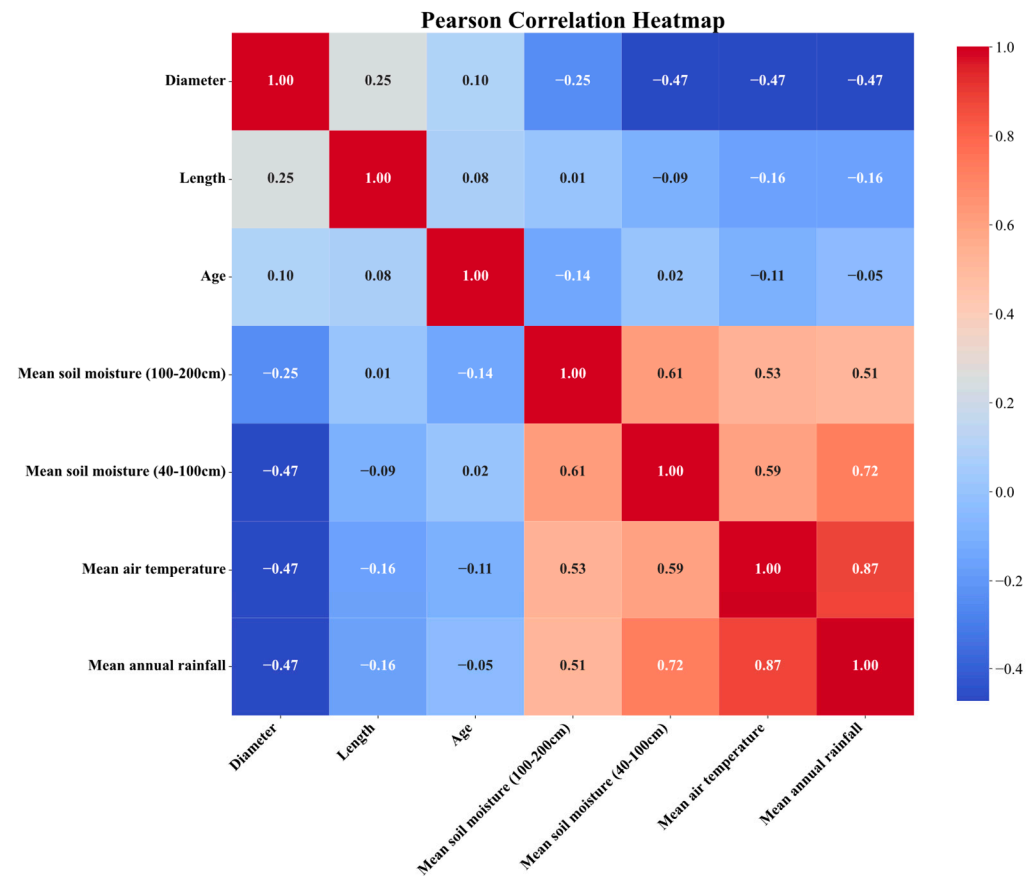


Figure 4. Pearson correlation heatmap.

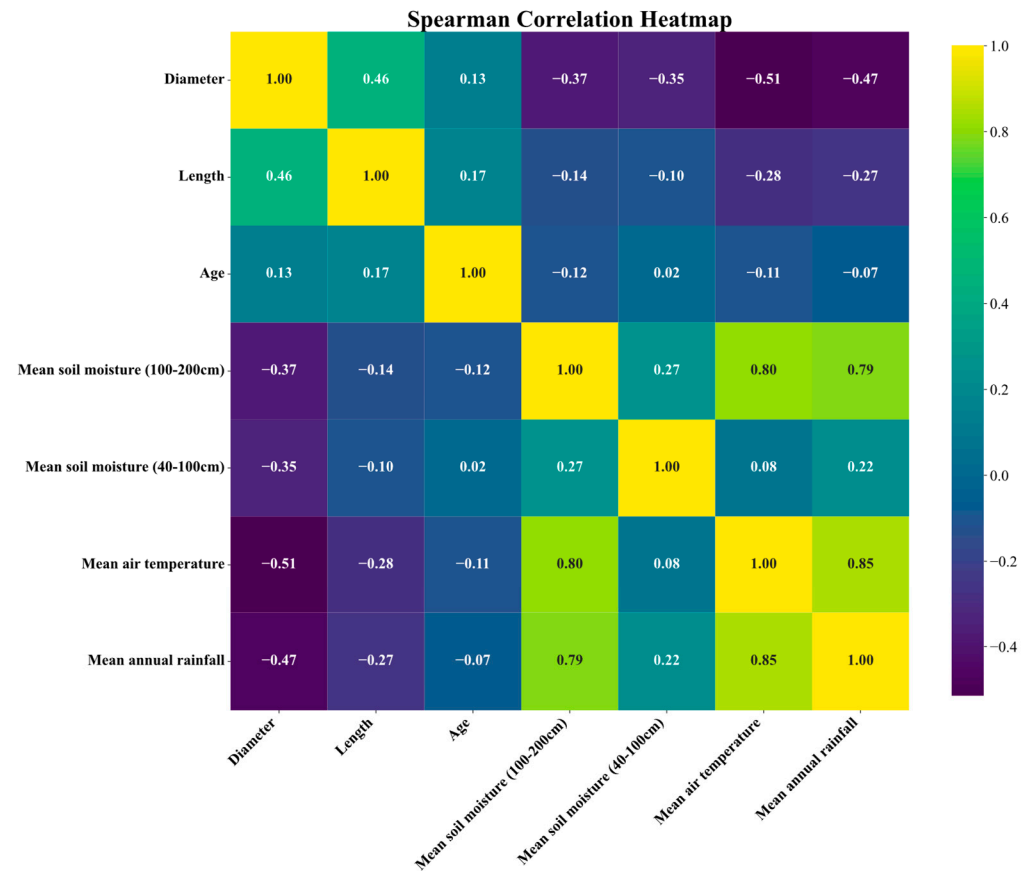


Figure 5. Spearman correlation heatmap.

As shown in the figures, all correlation coefficients among the selected features are below 0.9, indicating that no strong inter-feature correlations exist. This suggests that multicollinearity is not severe and that each feature can provide relatively independent information for model training. The results of the Pearson and Spearman analyses are also generally consistent, further confirming that the relationships among the variables are weak or weakly nonlinear. Overall, these findings demonstrate the rationality of the selected feature set and its suitability for subsequent modeling.

The target variable represents the pipeline condition status provided in the original Zenodo dataset [45], where each pipeline segment is pre-assigned to one of three condition categories: red, amber, or green. Following the original dataset specification, these categories were encoded as 0, 1, and 2, respectively, for model development, with red indicating poor pipeline condition, amber indicating moderate condition, and green indicating good condition. Since these condition labels were directly adopted from the publicly available dataset, no additional relabeling, recalculation, or modification was performed in this study. The class distribution is illustrated in Figure 6.

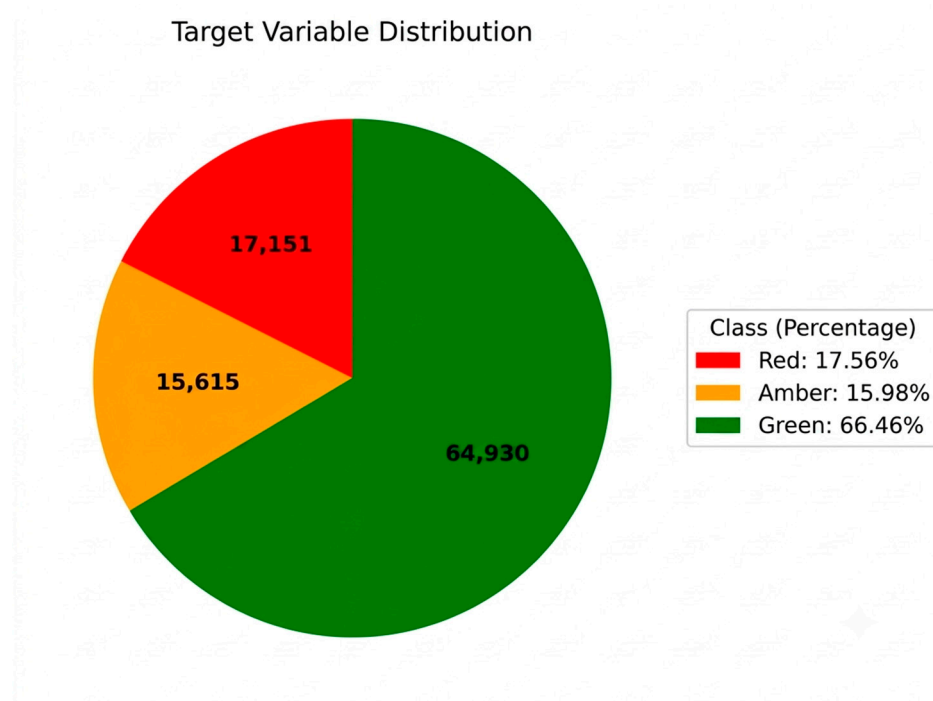


Figure 6. Target variable distribution.

As illustrated in Figure 6, the distribution of the target classes shows a noticeable imbalance. The green class accounts for 66.46% of the dataset, making it the dominant category. In contrast, the amber class and red class represent 15.98% and 17.56% of the samples, respectively. This skewed distribution indicates that most pipelines are in relatively good condition, while deteriorated or intermediate conditions are less frequently observed.

4. Results and Discussion

All models used in this study were implemented using the Scikit-learn library in Python 3.10, and the computations were conducted in the Jupyter Notebook 7.6 platform.

4.1. Model Training and Results

To ensure a rigorous and unbiased evaluation of model generalization, the final dataset was divided into training and testing subsets using a spatially constrained stratified

group sampling strategy rather than conventional random stratified sampling. Random splitting may introduce spatial data leakage because geographically adjacent pipelines or records from the same region can be assigned to both the training and testing sets, resulting in overly optimistic performance estimates. Therefore, pipeline geographic coordinates were first clustered using K-means, and the resulting spatial clusters were treated as grouping units during data partitioning. A stratified group splitting strategy was subsequently applied to generate an approximately 80:20 train–test split while preserving class distributions and ensuring that pipelines within the same spatial cluster were assigned exclusively to one subset. The independent test set was isolated throughout the entire modeling process and was used only for final performance evaluation. All preprocessing, model training, and hyperparameter optimization were performed using the training set. Model development was conducted using five-fold stratified cross-validation, where four folds were used for training and one fold for validation, with validation performance averaged across the five folds. Class weights were incorporated to alleviate class imbalance, with weights calculated only from the corresponding training folds and applied to the validation folds. Hyperparameter optimization was performed using the Optuna framework, and the optimal configurations were selected based on cross-validation performance.

For the proposed Stacking ensemble, three heterogeneous tree-based models, including RF, XGBoost, and LightGBM, were employed as base learners, while Logistic Regression was adopted as the meta-learner. These base learners were selected because they represent complementary ensemble learning paradigms with different modeling characteristics. RF provides robust bagging-based predictions with strong resistance to overfitting, whereas XGBoost and LightGBM employ gradient boosting strategies with distinct optimization mechanisms, enabling them to capture complex non-linear relationships from different perspectives. Their diversity enhances the complementarity of the generated meta-features and improves the effectiveness of the Stacking framework. A cross-validation-based procedure was implemented to generate unbiased meta-features. Specifically, the training set was divided into five folds, where base learners were trained using four folds and used to predict the remaining held-out fold. This process was repeated until each sample received an out-of-fold (OOF) probability prediction generated by a model that had not been trained on that sample. The OOF predictions from the three base learners were subsequently concatenated as input features for training the meta-learner. After hyperparameter optimization, the base learners were retrained using the complete training set with the selected parameters, and their predictions on the independent test set were integrated by the trained meta-learner to obtain the final Stacking classification results. CatBoost was additionally evaluated as an independent benchmark because it is a state-of-the-art gradient boosting algorithm with an alternative tree-growing strategy and ordered boosting mechanism, providing a strong reference for comparison while maintaining consistency in the optimization and evaluation protocol.

Table 2 presents the overall performance comparison of all models on the independent test dataset, and Figure 7 visualizes the corresponding evaluation metrics.

According to Table 2 and Figure 7, the Stacking ensemble framework achieves the highest values for Macro Precision, Macro Recall, Macro F1-score, and MCC. The agreement between Macro Precision and Macro Recall reflects a consistent trade-off between false positive control and sensitivity across all classes, which directly accounts for its elevated Macro F1-score and MCC. Among the individual baseline models, Random Forest presents higher overall metrics than LightGBM, XGBoost, and CatBoost, while also attaining the highest macro-averaged PR AUC among all evaluated models, indicating its capacity to rank positive instances effectively under class imbalance. These observations indicate that combining distinct base classifiers within a Stacking structure helps integrate

different model characteristics, leading to relatively more balanced overall prediction outcomes across the evaluated criteria.

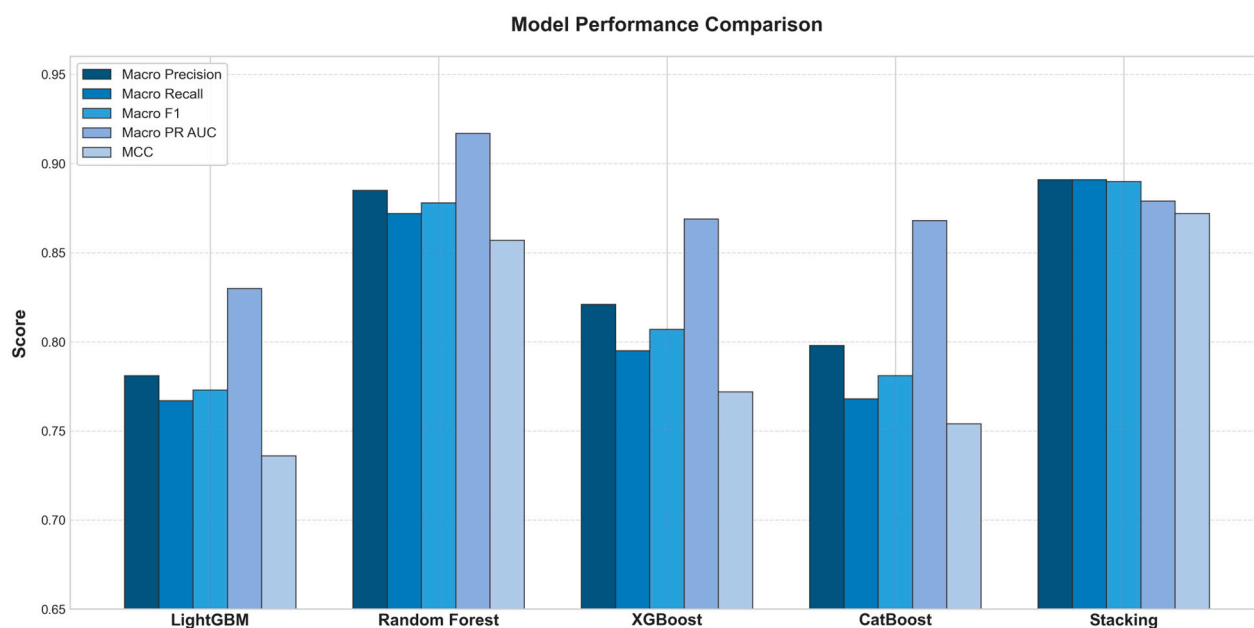


Figure 7. Model overall performance comparison.

Table 2. Overall performance comparison.

Model	Macro Precision	Macro Recall	Macro F1	Macro-Averaged PR AUC	MCC
LightGBM	0.781	0.767	0.773	0.830	0.736
Random Forest	0.885	0.872	0.878	0.917	0.857
XGBoost	0.821	0.795	0.807	0.869	0.772
CatBoost	0.798	0.768	0.781	0.868	0.754
Stacking	0.891	0.891	0.890	0.879	0.872

To further assess the operational viability and safety performance of the models, it is essential to look beyond macro-averaged metrics and examine class-specific performance for high-risk assets, like red pipes. In practical infrastructure management, misclassifying a severely degraded pipe as low risk entails significantly higher failure costs and safety hazards than false alarms. To quantify these operational risks, a cost loss metric is introduced with misclassification weights set to 10.0 for assigning red pipes to green status and 1.0 for all other error types. Table 3 presents the class-specific performance and cost loss metrics for the red class, Figure 8 visualizes these metric comparisons, and Figure 9 shows the confusion matrix for each model.

Table 3 reports the class-specific metrics and calculated cost loss for the red class across all evaluated models. The Stacking ensemble model obtains the highest Precision, Recall, and F1-score for this high-risk category, while also recording the lowest Cost Loss value among all configurations. Among the individual classifiers, Random Forest demonstrates competitive performance, achieving higher Precision and the highest PR AUC compared to the other baseline models. Other algorithms, including LightGBM, XGBoost, and CatBoost, result in higher Cost Loss values. These quantitative results indicate that the Stacking framework reduces severe misclassification instances for high-risk assets, thereby lowering overall operational risk costs.



Figure 8. Red class performance and cost loss comparison.

Table 3. Class-specific Performance Metrics for red Pipes.

Model	Precision	Recall	F1	PR AUC	Cost Loss
LightGBM	0.829	0.762	0.794	0.901	4247
Random Forest	0.902	0.847	0.873	0.945	3310
XGBoost	0.871	0.797	0.833	0.925	4126
CatBoost	0.816	0.804	0.810	0.919	4320
Stacking	0.919	0.850	0.883	0.911	2561

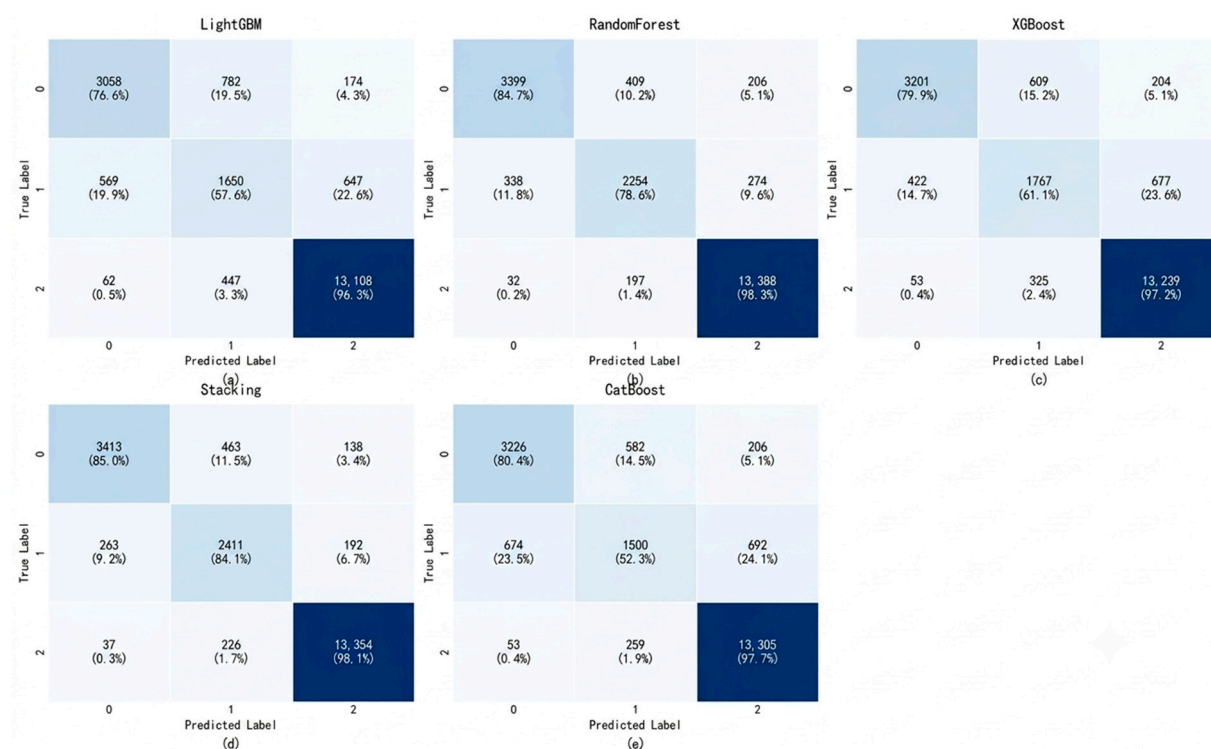


Figure 9. Confusion matrices for the evaluated models, including (a) LightGBM, (b) Random Forest, (c) XGBoost, (d) Stacking ensemble, and (e) CatBoost.

Figure 9 displays the confusion matrices of all models, illustrating the detailed sample distribution between true and predicted labels. Focus is placed on True Label 0, which corresponds to the high-risk red pipe class. The Stacking framework correctly identifies the largest number of Class 0 samples compared to all individual baseline models. More importantly, Stacking limits the critical error of misclassifying Class 0 as low-risk Class 2 to the lowest count among all evaluated models. This reduction in high-severity misclassification errors directly accounts for the lower Cost Loss observed in the Stacking ensemble.

A Copeland ranking analysis is further conducted to synthesize the results into a unified ordering. As a descriptive aggregation method, the Copeland score provides an intuitive summary of pairwise model comparisons across all evaluated criteria rather than direct evidence of statistical superiority. As shown in Figure 10, the Stacking ensemble receives the highest Copeland score, reflecting a more consistent overall balance across multiple indicators.

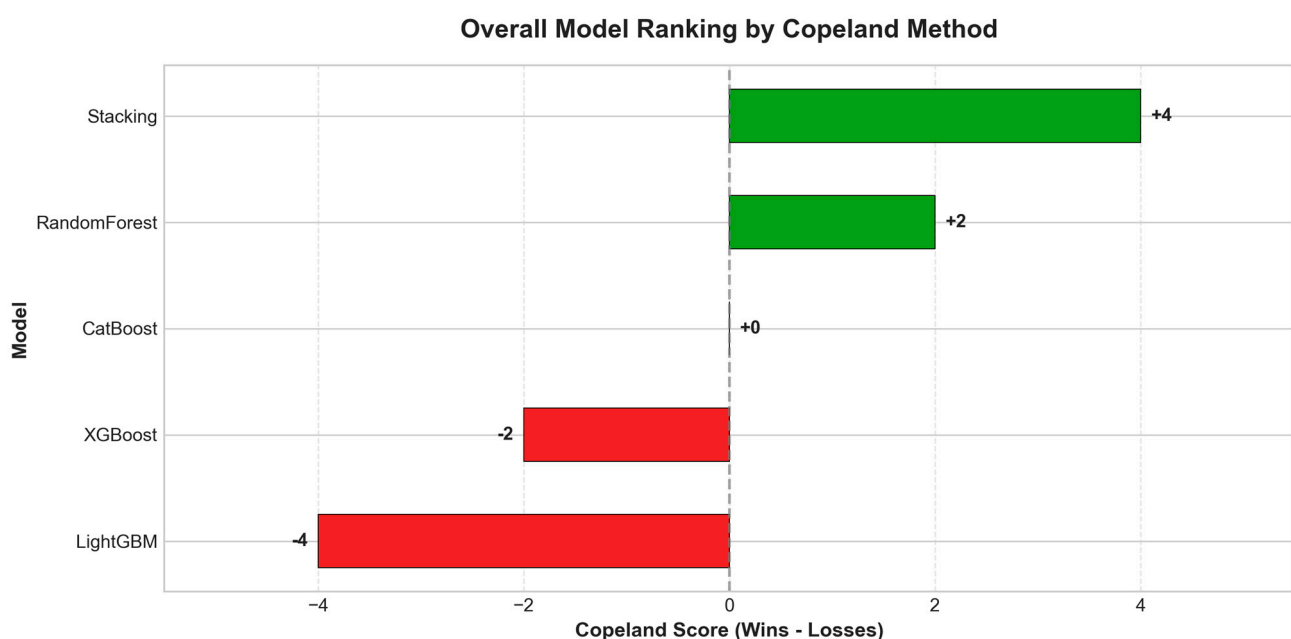


Figure 10. Copeland ranking of models.

4.2. Interpretability of Models

To enhance the interpretability of the models, we applied SHAP analysis to quantify feature contributions. Figures 11–14 show the results of the SHAP analysis.

To clarify the decision-making patterns of the model across different degradation states, class-specific SHAP analysis was conducted alongside overall feature importance evaluations. As summarized by the mean absolute SHAP values, pipe age, polyethylene (PE80) material, physical dimensions, including diameter and length, and environmental covariates such as mean air temperature and mean soil moisture demonstrate the highest relative contributions to the predictions across all condition classes.

A detailed examination of the class-specific beeswarm distributions reveals distinct feature-effect dynamics across individual risk levels. For severely degraded assets in the red class, encoded as 0, advanced pipe age exhibits predominantly positive SHAP values, showing a positive association with high-risk predictions. Pipe materials display pronounced conditional influence: while cast iron (CI) material exhibits positive SHAP values for specific instances, modern flexible materials such as medium-density polyethylene (MDPE), PE80, and unplasticized polyvinyl chloride (PVC) consistently yield negative SHAP values, indicating a lower likelihood of severe risk classification. Furthermore,

lower mean air temperatures and localized variations in deep soil moisture at depths of 100 to 200 cm contribute positively to severe risk predictions. In contrast, for intermediate risk states in the amber class, encoded as 1, predictive contributions shift toward a combination of age, structural dimensions, and material types; PE80 material displays negative SHAP values when present, whereas pipe length and diameter exhibit dispersed distributions that reflect non-linear contributions to moderate degradation estimates. Finally, for low-risk assets in the green class, encoded as 2, younger pipe age and materials such as high-density PE100 and unplasticized polyvinyl chloride (UPVC) yield positive SHAP values, shifting predictions toward the safe category, whereas elevated air temperatures and older age exert negative SHAP impacts on safe condition classification. Overall, these class-specific SHAP patterns illustrate how individual features influence predictions across distinct risk classes.

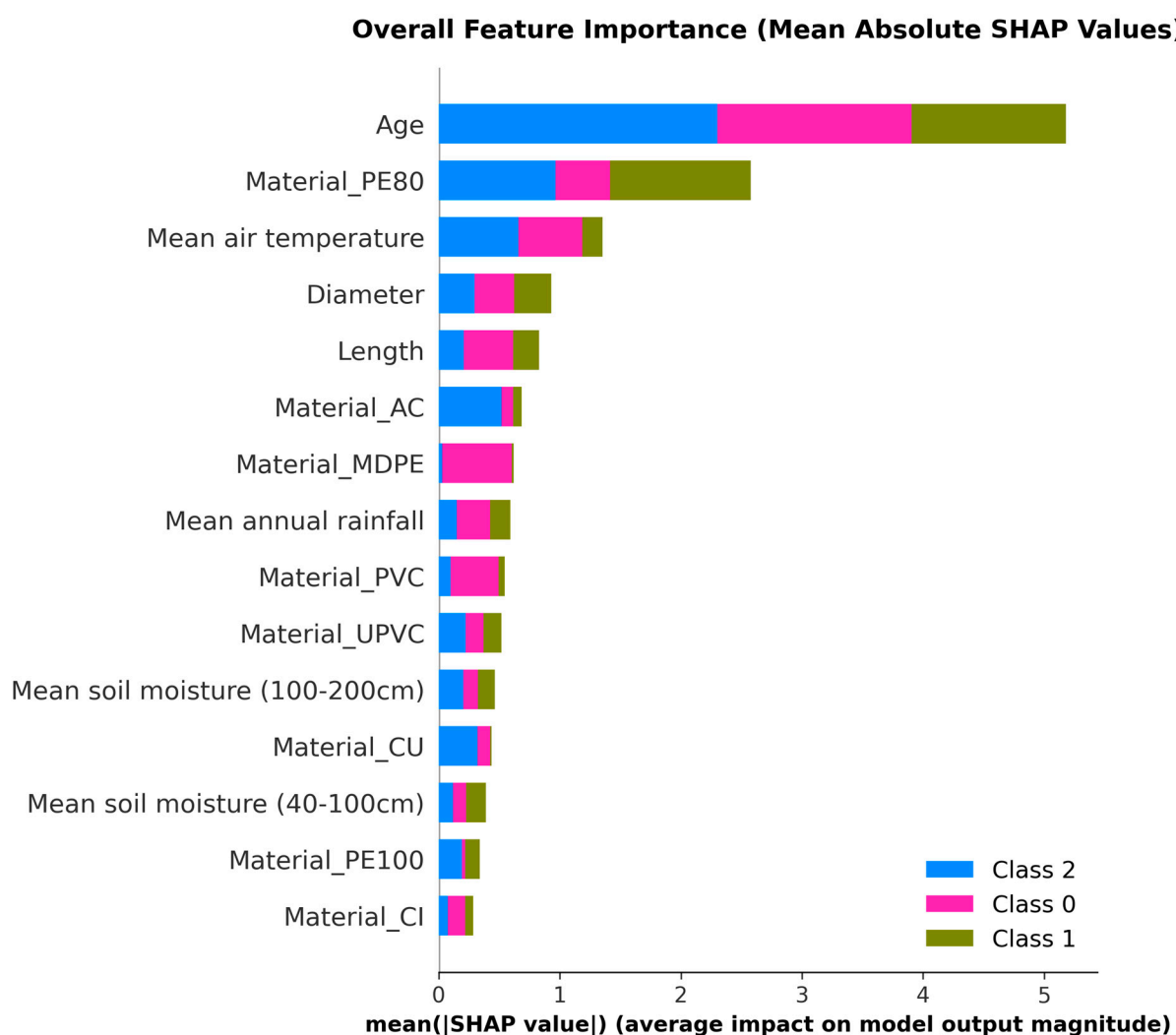


Figure 11. Overall feature importance.

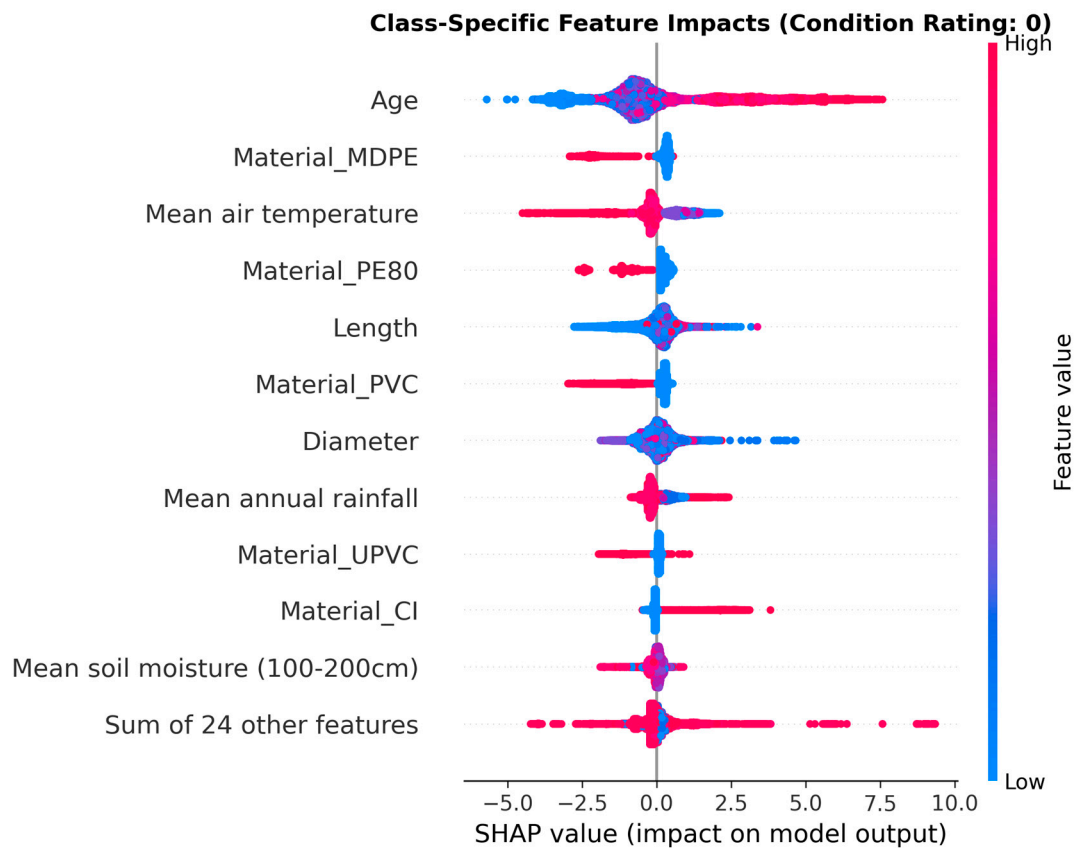


Figure 12. Class-specific feature impacts (red class).

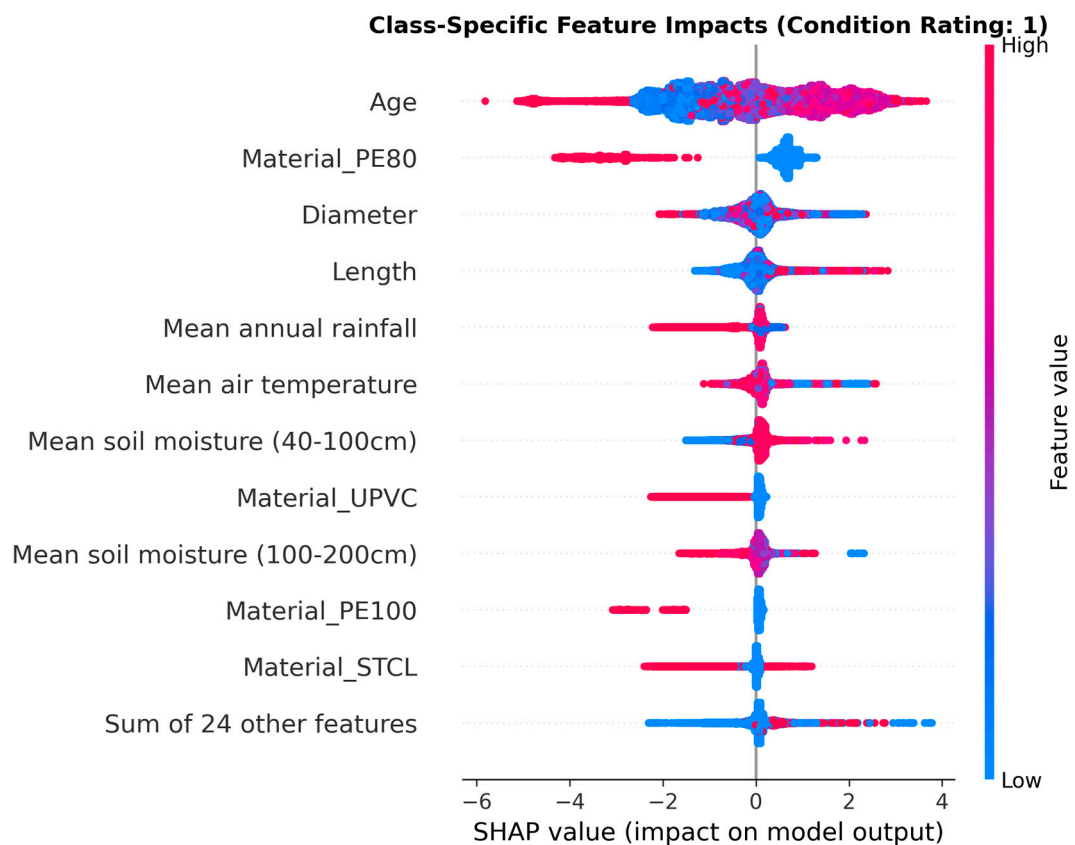


Figure 13. Class-specific feature impacts (amber class).

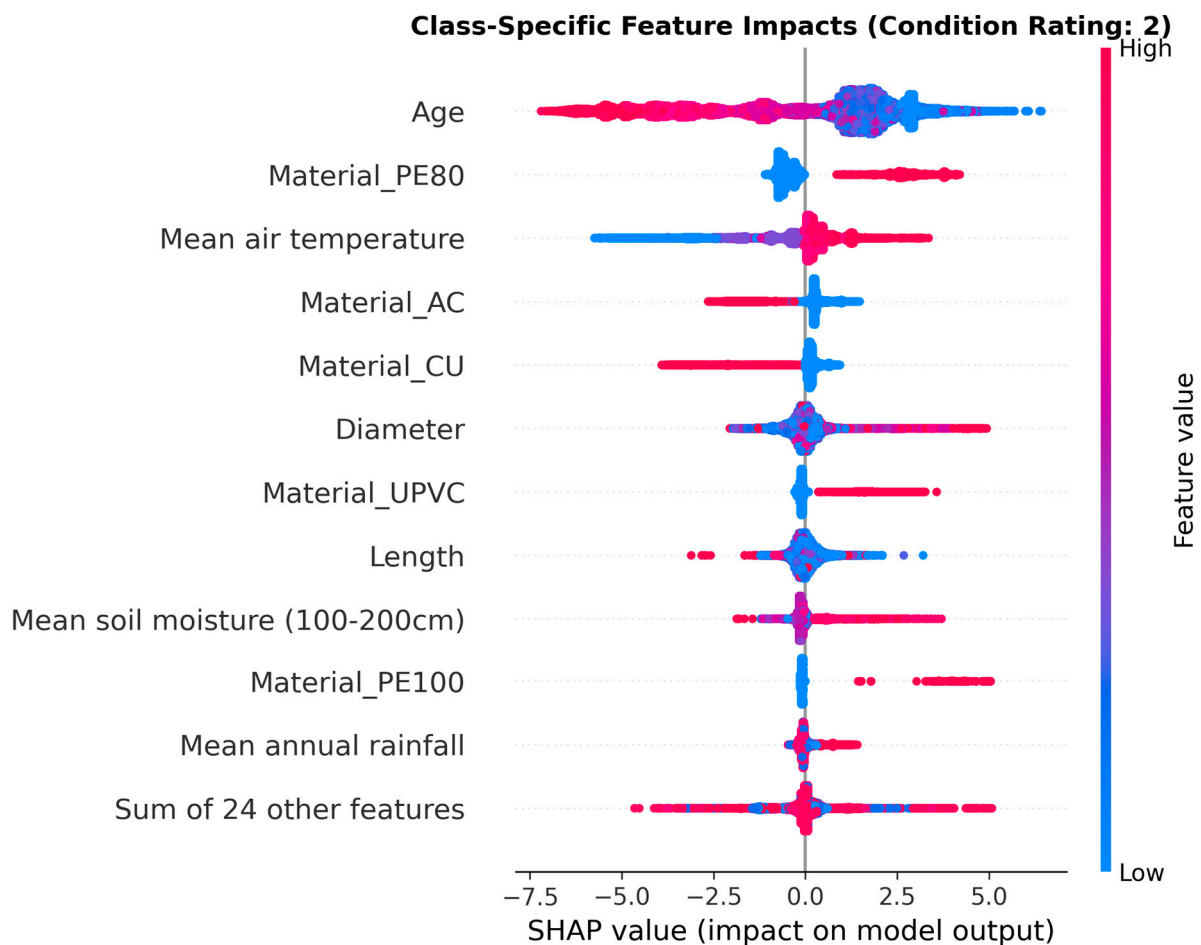


Figure 14. Class-specific feature impacts (green class).

5. Conclusions

In this study, a Stacking ensemble learning framework was developed for WDN condition prediction and pipeline failure risk assessment. By integrating tree-based base algorithms, including RF, XGBoost, and LightGBM with a Logistic Regression meta-learner, the framework captures non-linear failure patterns while mitigating class imbalance. Performance evaluations on the investigated network dataset indicate that the Stacking model achieved higher performance than the individual base learners as well as CatBoost, which was evaluated as an independent benchmark model. Specifically, the Stacking framework obtained higher Macro Precision, Macro Recall, Macro F1, and MCC scores than the baseline algorithms, while reducing false-negative classifications in the high-risk red category and yielding a lower operational Cost Loss. Class-specific SHAP analysis was conducted to interpret the model decision patterns across degradation states. Within the analyzed dataset, pipe age, material, physical dimensions, air temperature, and soil moisture demonstrated the highest relative contributions to the model outputs, with varying directional impacts across risk categories. In particular, advanced pipe age, CI material, lower air temperatures, and deep soil moisture were associated with higher risk contributions in the red class, whereas modern materials such as MDPE, PE80, and UPVC generally aligned with lower-risk estimations. Overall, integrating multi-model ensemble techniques with class-specific interpretability offers a structured, transparent analytical approach to support proactive asset management in water supply networks. From a practical perspective, the multi-tier Stacking architecture inherently introduces greater computational complexity and longer execution times than the individual base algorithms during the training phase. However, because pipeline condition assessment and risk pri-

oritization in municipal systems are typically conducted as periodic, offline tasks, the observed computational requirement remains within a manageable range for practical application. More importantly, performance evaluations suggest that this architecture enhances sensitivity toward high-risk pipeline segments, helping to mitigate false-negative predictions where severely degraded assets might otherwise be classified as low risk. In urban water infrastructure management, balancing a moderate increase in offline training time against improved accuracy in critical failure identification represents a reasonable practical trade-off.

Despite the observed performance of the proposed methodology, several limitations should be noted to contextualize the findings and guide future research. First, because the empirical evaluation relies on data from a single municipal water distribution system, the observed feature contribution patterns may reflect localized pipe installation standards, soil characteristics, and operational management conventions. Second, machine learning models are inherently constrained by the accuracy and completeness of historical asset condition registries, where unrecorded maintenance interventions, past pipe repairs, or localized physical stress events can introduce unmodeled noise into the dataset targets. Third, the current modeling scheme relies primarily on static inventory features and macro-level environmental descriptors, without directly linking prediction outputs to dynamic hydraulic parameters such as operational pressure fluctuations or transient flow velocities. Addressing these constraints in future work through cross-network transfer learning, integration of high-frequency hydraulic monitoring data, and automated risk prioritization algorithms will further improve model generalizability and practical utility for municipal water infrastructure management.

Author Contributions: Conceptualization, Q.L. and A.C.; methodology, A.C.; software, A.C.; validation, Q.L. and A.C.; formal analysis, A.C.; investigation, Q.L.; resources, Q.L.; data curation, Q.L.; writing—original draft preparation, A.C.; writing—review and editing, Q.L.; visualization, A.C.; supervision, Q.L.; project administration, Q.L.; funding acquisition, Q.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2022YFC3801000. The APC was also funded by the National Key Research and Development Program of China.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study. Requests to access the datasets should be directed to aochen@gs.zzu.edu.cn.

Acknowledgments: We thank Siyu Lu, Jiangchao Hu, and Xin Fu for their assistance with software development and data analysis.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Berardi, L.; Ugarelli, R.; Rostum, J.; Giustolisi, O. Assessing mechanical vulnerability in water distribution networks under multiple failures. *Water Resour. Res.* **2014**, *50*, 2586–2599. <https://doi.org/10.1002/2013wr014770>.
2. Housh, M.; Ostfeld, A. An integrated logit model for contamination event detection in water distribution systems. *Water Res.* **2015**, *75*, 210–223. <https://doi.org/10.1016/j.watres.2015.02.016>.
3. Kumpel, E.; Nelson, K.L. Intermittent Water Supply: Prevalence, Practice, and Microbial Water Quality. *Environ. Sci. Technol.* **2016**, *50*, 542–553. <https://doi.org/10.1021/acs.est.5b03973>.

4. Taylor, D.D.J.; Slocum, A.H.; Whittle, A.J. Demand Satisfaction as a Framework for Understanding Intermittent Water Supply Systems. *Water Resour. Res.* **2019**, *55*, 5217–5237. <https://doi.org/10.1029/2018WR024124>.
5. Fan, X.D.; Wang, X.W.; Zhang, X.J.; Yu, X. Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors. *Reliab. Eng. Syst. Saf.* **2022**, *219*, 108185. <https://doi.org/10.1016/j.res.2021.108185>.
6. Robles-Velasco, A.; Cortés, P.; Muñuzuri, J.; Onieva, L. Prediction of pipe failures in water supply networks using logistic regression and support vector classification. *Reliab. Eng. Syst. Saf.* **2020**, *196*, 106754. <https://doi.org/10.1016/j.res.2019.106754>.
7. Rajani, B.; Kleiner, Y. Comprehensive review of structural deterioration of water mains: Physically based models. *Urban Water* **2001**, *3*, 151–164. [https://doi.org/10.1016/S1462-0758\(01\)00032-2](https://doi.org/10.1016/S1462-0758(01)00032-2).
8. Kleiner, Y.; Rajani, B. Comprehensive review of structural deterioration of water mains: Statistical models. *Urban Water* **2001**, *3*, 131–150. [https://doi.org/10.1016/S1462-0758\(01\)00033-4](https://doi.org/10.1016/S1462-0758(01)00033-4).
9. Dawood, T.; Elwakil, E.; Novoa, H.M.; Gárate Delgado, J.F. Water pipe failure prediction and risk models: State-of-the-art review. *Can. J. Civ. Eng.* **2019**, *47*, 1117–1127. <https://doi.org/10.1139/cjce-2019-0481>.
10. Osman, H.; Bainbridge, K. Comparison of Statistical Deterioration Models for Water Distribution Networks. *J. Perform. Constr. Facil.* **2011**, *25*, 259–266. [https://doi.org/10.1061/\(ASCE\)CF.1943-5509.0000157](https://doi.org/10.1061/(ASCE)CF.1943-5509.0000157).
11. Farmani, R.; Kakoudakis, K.; Behzadian, K.; Butler, D. Pipe Failure Prediction in Water Distribution Systems Considering Static and Dynamic Factors. *Procedia Eng.* **2017**, *186*, 117–126. <https://doi.org/10.1016/j.proeng.2017.03.217>.
12. Kabir, G.; Tesfamariam, S.; Loepky, J.; Sadiq, R. Predicting water main failures: A Bayesian model updating approach. *Knowl.-Based Syst.* **2016**, *110*, 144–156. <https://doi.org/10.1016/j.knosys.2016.07.024>.
13. Ward, B.; Selby, A.; Gee, S.; Savic, D. Deterioration modelling of small-diameter water pipes under limited data availability. *Urban Water J.* **2017**, *14*, 743–749. <https://doi.org/10.1080/1573062X.2016.1254252>.
14. Geem, Z.W.; Tseng, C.-L.; Kim, J.; Bae, C. Trenchless Water Pipe Condition Assessment Using Artificial Neural Network. In *Proceedings of the Pipelines 2007: Advances and Experiences with Trenchless Pipeline Projects, Boston, MA, USA, 8–11 July 2007*; ASCE: Reston, VA, USA, 2012; pp. 1–9.
15. Elshaboury, N.; Marzouk, M. Comparing machine learning models for predicting water pipelines condition. In *Proceedings of the 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 24–26 October 2020*; IEEE: New York, NY, USA, 2020; pp. 134–139.
16. Fan, X.; Zhang, X.; Yu, X.B. Uncertainty quantification of a deep learning model for failure rate prediction of water distribution networks. *Reliab. Eng. Syst. Saf.* **2023**, *236*, 109088. <https://doi.org/10.1016/j.res.2023.109088>.
17. Jafari, S.M.; Zahiri, A.R.; Bozorg Hadad, O.; Mohammad Rezapour Tabari, M. A hybrid of six soft models based on ANFIS for pipe failure rate forecasting and uncertainty analysis: A case study of Gorgan city water distribution network. *Soft Comput.* **2021**, *25*, 7459–7478. <https://doi.org/10.1007/s00500-021-05706-4>.
18. Zali, R.B.; Latifi, M.; Javadi, A.A.; Farmani, R. Semisupervised Clustering Approach for Pipe Failure Prediction with Imbalanced Data Set. *J. Water Resour. Plan. Manag.* **2024**, *150*, 04023078. <https://doi.org/10.1061/JWRMD5.WRENG-6263>.
19. Liu, W.; Xie, Z.; Song, Z. Predicting Water Pipe Failures Using Deep Learning Algorithms. *J. Infrastruct. Syst.* **2023**, *29*, 04023022. <https://doi.org/10.1061/JITSE4.ISENG-2247>.
20. Robles-Velasco, A.; Cortés, P.; Muñuzuri, J.; De Baets, B. Prediction of pipe failures in water supply networks for longer time periods through multi-label classification. *Expert Syst. Appl.* **2023**, *213*, 119050. <https://doi.org/10.1016/j.eswa.2022.119050>.
21. Cen, H.; Huang, D.; Liu, Q.; Zong, Z.; Tang, A. Application Research on Risk Assessment of Municipal Pipeline Network Based on Random Forest Machine Learning Algorithm. *Water* **2023**, *15*, 1964. <https://doi.org/10.3390/w15101964>.
22. Tavakoli, R.; Sharifara, A.; Najafi, M. Artificial Neural Networks and Adaptive Neuro-Fuzzy Models to Predict Remaining Useful Life of Water Pipelines. In *Proceedings of the World Environmental and Water Resources Congress 2020, Henderson, NV, USA, 17–21 May 2020*; ASCE: Reston, VA, USA, 2020; pp. 191–204.
23. Hu, Q.; Che, D.; Zhang, Q.; Zhou, J.; Wang, F.; Zhang, Z.; Song, Z. Improving underground pipeline resilience: Prediction and interpretability analysis of urban water distribution network pipe failures during cold waves using machine learning. *Tunn. Undergr. Space Technol.* **2025**, *163*, 106717. <https://doi.org/10.1016/j.tust.2025.106717>.
24. Bahrami Chegeni, I.; Riyahi, M.M.; Bakhshipour, A.E.; Azizipour, M.; Haghighi, A. Developing Machine Learning Models for Optimal Design of Water Distribution Networks Using Graph Theory-Based Features. *Water* **2025**, *17*, 1654. <https://doi.org/10.3390/w17111654>.
25. Rahmanshahi, M.; Duan, H.-F.; Keramat, A.; Rad, N.V.; Nadian, H.A. Assessing Viscoelastic Parameters of Polymer Pipes via Transient Signals and Artificial Neural Networks. *Eng. Proc.* **2024**, *69*, 74. <https://doi.org/10.3390/engproc2024069074>.

26. Sharififard, E.; Azizpour, M.; Ahadiyan, J.; Haghighi, A. Determination of creep function coefficients of viscoelastic pipes using a transient-guided machine learning model. *AQUA-Water Infrastruct. Ecosyst. Soc.* **2024**, *73*, 2132–2149. <https://doi.org/10.2166/aqua.2024.091>.
27. Vittori, G.; Falkouskaya, Y.; Jimenez-Gutierrez, D.M.; Cattai, T.; Chatzigiannakis, I. Graph neural networks to model and optimize the operation of Water Distribution Networks: A review. *J. Ind. Inf. Integr.* **2025**, *47*, 100880. <https://doi.org/10.1016/j.jii.2025.100880>.
28. Candelieri, A.; Giordani, I.; Archetti, F. Supporting Resilience Management of Water Distribution Networks through Network Analysis and Hydraulic Simulation. In *Proceedings of the 2017 21st International Conference on Control Systems and Computer Science (CSCS), Bucharest, Romania, 29–31 May 2017*; IEEE: New York, NY, USA, 2017; pp. 599–605.
29. Candelieri, A.; Soldi, D.; Archetti, F. Cost-effective sensors placement and leak localization—the Neptun pilot of the ICeWater project. *J. Water Supply Res. Technol.-Aqua* **2015**, *64*, 567–582. <https://doi.org/10.2166/aqua.2015.037>.
30. Ponti, A.; Candelieri, A.; Giordani, I.; Archetti, F. A Novel Graph-Based Vulnerability Metric in Urban Network Infrastructures: The Case of Water Distribution Networks. *Water* **2021**, *13*, 1502. <https://doi.org/10.3390/w13111502>.
31. Fan, Z.; Yu, Z.; Yang, K.; Chen, W.; Liu, X.; Li, G.; Yang, X.; Chen, C.L.P. Diverse Models, United Goal: A Comprehensive Survey of Ensemble Learning. *CAAI Trans. Intell. Technol.* **2025**, *10*, 959–982. <https://doi.org/10.1049/cit.2.70030>.
32. Yu, Z.; Wang, D.; Zhao, Z.; Chen, C.L.P.; You, J.; Wong, H.S.; Zhang, J. Hybrid Incremental Ensemble Learning for Noisy Real-World Data Classification. *IEEE Trans. Cybern.* **2019**, *49*, 403–416. <https://doi.org/10.1109/TCYB.2017.2774266>.
33. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. <https://doi.org/10.1007/BF00058655>.
34. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; ACM: New York, NY, USA, 2016; pp. 785–794.
35. Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, Bari, Italy, 3–6 July 1996*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1996; pp. 148–156.
36. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. <https://doi.org/10.1214/aos/1013203451>.
37. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).
38. Hou, H.; Liu, C.; Wei, R.; He, H.; Wang, L.; Li, W. Outage duration prediction under typhoon disaster with stacking ensemble learning. *Reliab. Eng. Syst. Saf.* **2023**, *237*, 109398. <https://doi.org/10.1016/j.res.2023.109398>.
39. Mao, Y.; Yu, X.; Wang, F.; Zhu, J. Electric vehicle charging demand forecasting: A data-driven integrated learning approach. *Renew. Energy* **2026**, *256*, 124141. <https://doi.org/10.1016/j.renene.2025.124141>.
40. Pang, M.; Zhang, Z.; Zhou, Z.; Zhou, W.; Li, Q. Estimated ultimate recovery prediction of shale gas wells based on stacked integrated learning algorithm. *Sci. Rep.* **2025**, *15*, 19258. <https://doi.org/10.1038/s41598-025-03801-2>.
41. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. <https://doi.org/10.1023/A:1010933404324>.
42. Furxhi, I.; Murphy, F.; Mullins, M.; Poland, C.A. Machine learning prediction of nanoparticle in vitro toxicity: A comparative study of classifiers and ensemble-classifiers using the Copeland Index. *Toxicol. Lett.* **2019**, *312*, 157–166. <https://doi.org/10.1016/j.toxlet.2019.05.016>.
43. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. <https://doi.org/10.1038/s42256-019-0138-9>.
44. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017*; pp. 4768–4777.
45. Thomas, M.; Dorian, B.; Iguchi, T.; Chini, C. *Code and Data for “Comparing Methods for Water Distribution System Condition Assessment and Forecasting”*; Zenodo: Geneva, Switzerland, 2025.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.