Journal of Infrastructure Preservation and Resilience

RESEARCH

Open Access

Inspection prioritization of gravity sanitary sewer systems using supervised machine learning algorithms

Karthikeyan Loganathan[1], Mohammad Najafi[1], Sharareh Kermanshachi[1*], Praveen Kumar Maduri[2] and Apurva Pamidimukkala[1]

Abstract

Underground wastewater collection systems degrade with time, necessitating utility owners to engage in ongoing evaluations and enhancements of their asset management frameworks to preserve the performance of their assets. The inspection and condition assessment of sewer pipes are crucial for the effective operation and maintenance of sewer systems. The closed-circuit television (CCTV) is frequently employed to examine sewer pipes in the United States. This procedure is both costly and laborious because of the extensive number of pipes in a metropolis. Prioritisation of inspection for sanitary sewage pipe segments requiring repair or maintenance can be done in advance depending on their past performance. Hence, the aim of this study is to construct a predictive model for the state of sanitary sewer pipes, utilising data collected from a city located in the southcentral region of the United States. The main contribution is that this study used multiclass classification and predicted PACP scores of the pipes. Condition prediction models were developed using extensively utilised supervised machine learning algorithms including logistic regression (LR), k-nearest neighbors (k-NN), and random forest (RF). However, the bulk of the constructed models were assessed using a limited number of assessment measures, such as the receiver operator characteristic (ROC) curve and the area under the curve (AUC) value. This paper asserts that the assessment of the predictive capacity of these models cannot be determined only by relying on ROC and AUC values. Out of the three models evaluated in this study, the LR model had an AUC value of 0.76. However, this model had a higher number of misclassifications or inaccurate predictions compared to the other models. Consequently, these models were assessed using additional assessment measures, including precision, recall, and F-1 scores (which represent the harmonic mean of precision and recall). Curiously, the LR model achieved an F1-score of 0.28 on a scale ranging from 0 to 1. The RF model yielded an F1-score of 0.45 and an AUC value of 0.86. The existing model can be enhanced before it is employed by asset managers during the inspection phase to assess the state of their sanitary sewers and identify essential sewers that require immediate care.

Keywords Inspection, Deterioration, Condition Prediction, Machine learning, Algorithm

Introduction

The underground pipeline networks in the United States span vast distance and constitute a substantial percentage of the wastewater infrastructure resources (Najafi and Gokhale [21]). Considering that the significant amounts of U.S. wastewater infrastructures are approximately 60 years old, any major breakdown of these systems might have a significant and disruptive impact on

*Correspondence:
Sharareh Kermanshachi
Sharareh.kermanshachi@uta.edu
[1] Department of Civil Engineering, University of Texas at Arlington, Arlington, TX 76019, USA
[2] Dean Academics, Galgotias College of Engineering and Technology, Noida 201306, India

the surrounding areas regarding their economy, society, and environment (EPA [6] and EPA [7]). Furthermore, the remediation of malfunctioning sanitary sewers can impose a substantial financial burden on governments, as well as affect public health (Najafi and Gokhale [20]).

Based on a research conducted by the Environmental Protection Agency (EPA), several municipalities have sewers that are over 100 years old (EPA [7]). Over a period of time, underground infrastructure degrade, requiring utility owners to undertake continual changes and establlish asset management frameworks to preserve the performance of their infrastructure (Najafi and Kulandaivel [22]). "Asset management is defined as managing infrastructure capital assets to minimize the total cost of owning and operating them, while delivering the service levels customers desire (Najafi and Gokhale [21])." The primary duties of an asset management plan involve examining and evaluating the state of an asset (Tscheikner-Gratl et al. [26]).

In the past, municipalities were responsible for handling the architecture, building, and administration of sanitary systems (Wirahadikusumah et al. [28]). The Pipeline Assessment Certification Programme (PACP), designed by the National Association of Sewer Service Companies (NASSCO) in 2002, is widely used in the US to evaluate sanitary sewers. The PACP assigns distinct codes to every potential flaw in sewer systems, along with a rating from 1 to 5 that indicates the structural soundness of each pipe segment. Figure 1 depicts the arrangement of the PACP inspection equipment. A NASSCO-certified operator carefully reviews the recorded footage and meticulously adds problem codes to either computer program or a spreadsheet.

The PACP is preconfigured with predetermined scores for each category of fault and their corresponding level of seriousness (NASSCO 2018). The final score of the sewer pipeline can be estimated according to the information as presented in Table 1.

Nevertheless, the examination of a sewer pipe is a costly and laborious procedure due to the recommendation of the PACP to limit the camera speed to a maximum of 30 feet or 9 m per minute. Consequently, because to the extensive number of assets that municipalities possess, it is not financially viable to conduct inspections on every individual sanitary sewage pipe (Malek Mohammadi [16]). Furthermore, towns have the option to check sewer lines that are in good structural condition. Otherwise, the significant budget allocated for this purpose

Table 1 PACP condition rating (Loganathan [14])

PACP	Description	Predicted Failure Time
1	Excellent	Highly improbable to experience failure in the near future
2	Good	20 years or more
3	Fair	10–20 years
4	Poor	5–10 years
5	Requires urgent attention	Failed or imminent failure within the next five years

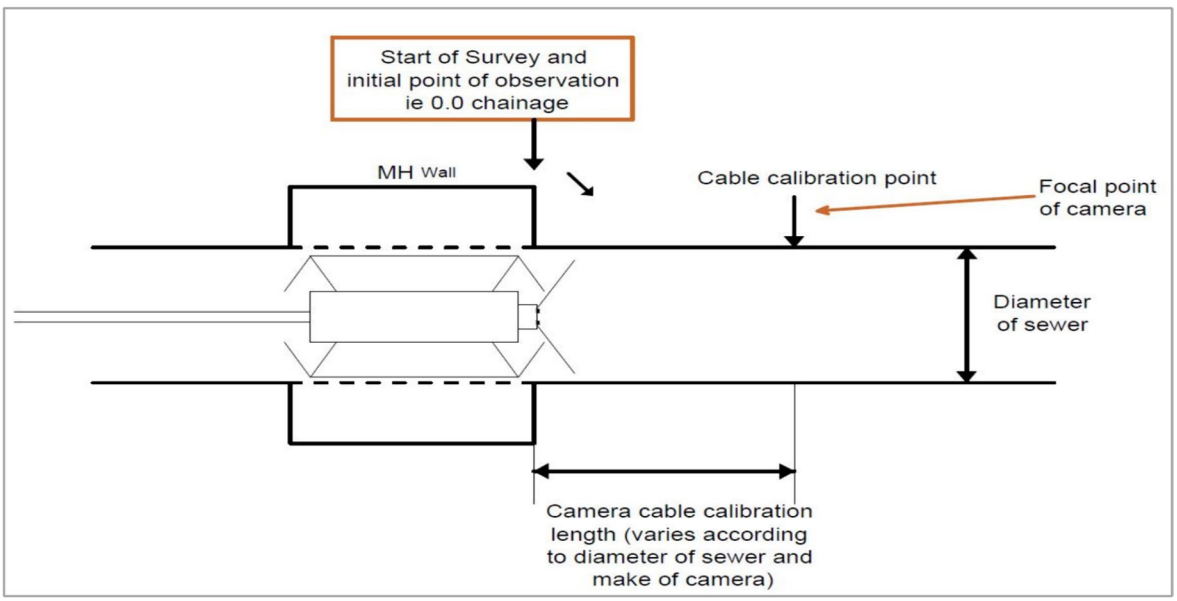

Fig. 1 PACP inspection equipment setup (Loganathan [14])

might be directed towards repairing and renewing segments of sewer pipes that are in need. To accomplish efficient budget allocation, one could predict the state of pipelines by analysing their past performance (Loganathan [14]). Therefore, accurately forecasting the state of sanitary sewer pipes would significantly advantage utility owners throughout the evaluation stages of condition assessment.

Condition prediction of sanitary sewer pipes

Evidently, not every sewer line in an registry would be functionaly deficient or be on the verge of breakdown. The examination of sewage pipelines can be restricted by closely examining the pipelines that are in poor state by anticipating their deterioration beforehand (Wright et al. [29]). The anticipation of the state of a sanitary sewer line is not a novel idea. Ariaratnam et al. in [4] created a binary classification model to forecast the proabability of a sewage infrastructure system being in a structurally inadequate condition. Hahn et al. in [10] built a sophisticated knowledge-based assistance system to prioritise sewer pipe inspection.

In 2005, Najafi and Kulandaivel constructed a predictive model for conditions utilising the Artificial Neural Network (ANN) approach. The model exhibited satisfactory performance throughout the training phase, however its performance was deemed unsatisfactory during the testing phase. Syacharni et al. ([25]) constructed a deterioration model for sewer pipes using a decision-tree approach. The study utilised a range of methodologies including regression analysis, neural networks and decision trees.

Harvey and McBean ([11]) established a predictive model to assess the structural quality of sewer pipelines. The random forests technique, which is a form of supervised machine learning, was employed to train the model. Hernandez et al. ([12]) constructed a predictive model for structural condition utilising diverse machine learning algorithms. The study conducted a comparative analysis of the performances exhibited by different models. The work conducted by Malek Mohammadi et al. ([17]) constructed predictive models for the quality of sanitary sewer lines using a range of machine learning methods. While the model demonstrated satisfactory accuracy, it categorised the condition ratings of pipes into binary classes instead of multiple condition ratings.

Various studies were performed to develop sanitary sewers condition prediction models using algorithms such as decision trees, k-nearest neighbors (k-NN), and so on (Ana et al. [3], Vladeanu [27]). Figure 2 presents the distribution of various machine learning techniques employed in considered studies. However, most of the developed models were either based on binary classification or validated based on more general metrics such as area under the curve (AUC) value and receiver operator characteristic (ROC) curve (Loganathan [14]). The main contribution is that this study used multiclass classification and predicted PACP scores of the pipes.

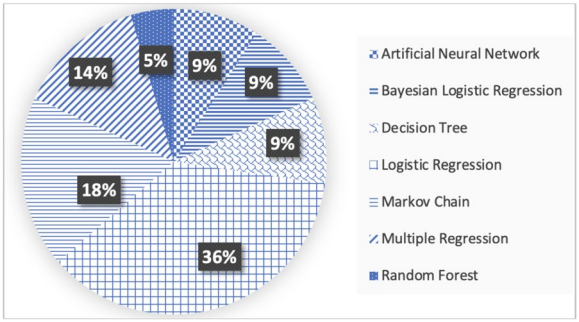

Fig. 2 ML techniques used in various research studies

Methodology

Typically, condition prediction models are created using supervised machine learning methods. Machine learning (ML) is a broad concept that encompasses computational algorithms that utilise previous data to provide accurate predictions (Mohri et al. [18]).

Machine learning can be categorised into two main types: unsupervised learning and supervised learning. The majority of data analysis conducted in various research pertaining to condition prediction is categorised as supervised learning. This involves training a computer program to analyse past data that contains the output variable. Using the knowledge gained from the training, a forecast is made for a fresh collection of data or data that has not been previously recorded. The predominant strategies utilised in ML models are typically derived from classification, regression, or a fusion of the two. Algorithms can be classified by their underlying working principles, such as k-Nearest Neighbours (a clustering method), and decision trees. Classification and regression are widely recognised as the primary divisions of supervised learning algorithms. The regression method is appropriate when there is a need to forecast a continuous dependent variable using several independent variables (Müller and Guido [19]). The study utilises a categorical dependent variable with 5 distinct classes, necessitating the employment of a classification machine learning technique to construct prediction models.

Data collection

Closed Circuit Telivsions (CCTVs) are extensively Utilized in the US to inspect sanitary sewage systems (NASSCO [23]). This work utilises historical data

obtained from a city in the southcentral region of the United States to create a predictive model for assessing the quality of sanitary sewer pipes. The model may be used to determine which pipes should be inspected first in future inspections. This study focuses exclusively on gravity flow sanitary sewage pipes, specifically ignoring force main systems. The sewage system inventory is recorded in geographic information system (GIS) databases, which contain detailed information about pipe installations, surrounding soil types, pipe placements on geographical maps, and other relevant data. The dataset collected consisted of 32,854 distinct pipe segment details. Table 2 displays a representative sample of the gathered data.

Primary data analysis

GIS_ID is an exclusive identifying code assigned by the operator during the inspection process for referencing purposes.

INSTALL_DATE is the specific date when the pipe was installed for use.

INSPEC_DATE is the specific date on which the inspection was finished. The disparity between the date of installation and the date of inspection would provide a crucial attribute, namely the sewage pipeline age. Although the initial inspection data is from 2000, only 500 sewer pipes were evaluated for condition till 2005.

INSPEC_LENGTH refers to the cumulative distance, measured in feet, that is examined from the starting manhole to the ending manhole of a certain section of a sanitary sewer. The inspection data indicate that the bulk of the sewer lines had a maximum length of 1,000 feet.

MAPSCOGRID serves as a geographic reference for the examined pipe segment. Since 1952, urban maps have been created using numbered grid systems and were commonly known as Mapscogrids.

DOWNELEV and UPELEV represent the heights, measured in feet above sea level, at the manholes located downstream, and upstream, respectively. This is crucial for determining the slope of the sewer line.

SUBAREA refers to the specific drainage basin in which the examined sewer line is situated. The subarea surrounding the sewage lines was identified using unique alphanumeric IDs based on the drainage basin. The initial two letters in the term SUBAREA indicate the specific basin type, including Village Creek (VC), Big Fossil (BF), Clear Fork (CF), and others.

The DIAMETER indicates the pipe's size or diameter, measured in inches. The sewer lines had a diameter that varied from 4 to 96 inches. It was observed that pipes with a diameter over 60 inches did not possesses PACP score of 5. This suggests that pipes with bigger diameters were in relatively good structural condition in comparision to pipes with smaller diameters.

The "MATERIAL" indicates the specific material utilised in the production of the sewer pipe. Most of the sewage lines are made of Poly Vinyl Chloride (PVC), with the most of them being built after 1980. The second greatest share consists of Vitrified Clay Pipes (VCP), which were primarily installed before 1980.

The collected data includes a PACP column that provides the PACP ratings for each length of the sewer line. The PACP scores range from 1 to 5, with 1 indicating a functionally good condition and 5 indicating a near failure condition, as previously mentioned.

Exploratory data analysis

The raw data obtained from the inventory of databases from one of cities in Texas were processed to utilize them for model development. The resulting dataset underwent processing and consisted of 32,751 pipe segments. This dataset was used for further research and included 7 independent factors and multi-class categorical

Table 2 Sample of data collected for the study

GIS_ID	INSPEC_ DATE	INSTALL_ DATE	MATERIAL	INSPEC_ LENGTH	MAPSCO GRID	DOWN ELEV	UP-ELEV	SUBAREA	DIAME-TER	PACP
60,718	12/12/2010	8/18/1988	CONCRETE	844	93G	550.88	551.3	VC09_01	54	2
60,717	6/6/2011	7/25/1958	CI	460	93G	551.25	551.44	VC08_01	39	2
60,723	11/16/2012	12/3/1964	VCP	259	89 F	668.57	673.9	CF05_03	6	3
60,724	11/19/2012	12/9/1964	CONCRETE	503	89 F	673.9	689.06	CF05_03	6	2
60,732	11/13/2014	6/11/1947	VCP	203	47Y	716.86	723.7	MC03_06	6	2
60,728	8/9/2017	4/25/2002	PVC	444	106U	607.88	610.87	VC11_01	24	1
60,719	12/28/2017	7/17/2001	PVC	415	46 L	790.35	792	MC04_04	8	1
60,729	4/28/2017	2/28/2002	PVC	396	106 S	645.67	647.17	VC11_01	24	2
60,720	1/5/2018	7/1/2004	PVC	426	46 H	752.45	760.1	MC04_04	8	1
60,726	8/7/2019	3/1/2005	PVC	112	119G	640.73	641.37	VC11_03	8	1

Table 3 Details of extracted features

Variable Type	Features Extracted	Description (Data Types)
Independent (Response Variables)	Age	Continuous Numerical
	Length	
	Slope	
	Diameter	
	MAPSCOGRID	Nominal
	Material	Nominal
	SUBAREA	Nominal
Dependent	PACP	Multi-class Categorical

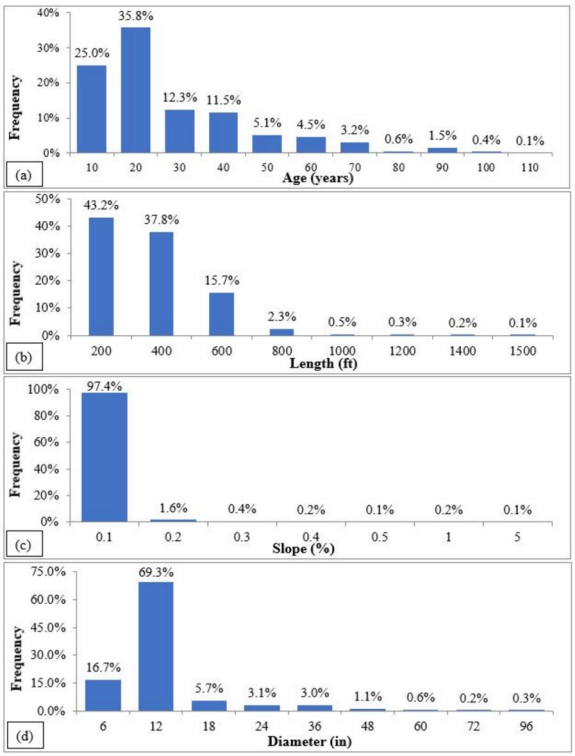

Fig. 3 Frequency distribution based on pipe: (**a**) Age, (**b**) Length, (**c**) Slope and (**d**) Diameter

dependent variables. Table 3 displays the specific information for each feature.

The initial data analysis involves extracting the age of the examined sewage pipeline segment. Figure 3 displays the distribution of independent variables. As depicted in Fig. 3(a), the age ranged up to a maximum of 107 years. As depicted in Fig. 3(b), the examined sewer pipe segment had a range of lengths, from 8 feet to 1,500 feet. Approximately 81% of the pipe segments were less than 400 feet. Notably, several observations were discovered to extend across a distance of more than 1,000 feet. Figure 3(c) demonstrates that the majority of the pipes, approximately 99%, were very flat with a maximum slope of 0.2%. The diameter distribution is depicted in Fig. 3(d), with over 90% of pipes having a diameter of less than 24 inches, while around 1% of pipes had a diameter exceeding 60 inches.

The distribution of pipe materials and their accompanying PACP scores is depicted in Fig. 4. It was discovered that PVC makes up a significant proportion of the sewer pipes, accounting for approximately 60%. This is followed by VCP and concrete pipes, which account for 17% and 9% respectively. Due to the prevalence of structurally compromised pipes across various pipe materials, the model development encompasses all sorts of pipe materials. Notably, a substantial quantity of concrete pipes were experiencing degradation, as depicted in Fig. 4. Table 4 displays a representative sample of the processed final dataset.

Figure 5 displays the distribution of the dependent variable, which is the PACP scores of individual pipes, for the condition prediction model. Approximately 90% of the pipelines are in a good structural condition, as indicated by PACP values of 1 and 2. Furthermore, it is worth mentioning that over 70% of the pipes have a lifespan of less than 30 years.

Model development

Python programming language, is employed in this work for the purpose of constructing prediction models. Python is chosen due to its open-source nature and its ability to efficiently handle extensive data libraries. The study utilised different Python libraries, which are listed in Table 5 along with their respective functions. This study utilized three machine learning models Logistic Regression (LR) as it provides a simple and interpretable model, k-Nearest Neighbours (k-NN) as it requires no assumptions about data distribution, and Random Forest (RF) as it offers high accuracy and robustness by combining multiple decision trees. This study did not consider support vector machine as they are well documented in the literature as the effective tools for binary classification (Burges, 1998; Shmilovici [24]). By utilising the specified libraries, machine learning methods like as LR, k-NN, and RF are trained using preprocessed data in order to create models for predicting conditions.

Logistic regression

Logistic regression (LR) is a frequently used statistical technique in the field of machine learning. The logistic regression approach is employed to examine the correlation between several independent factors and a categorical dependent variable. This approach involves fitting the

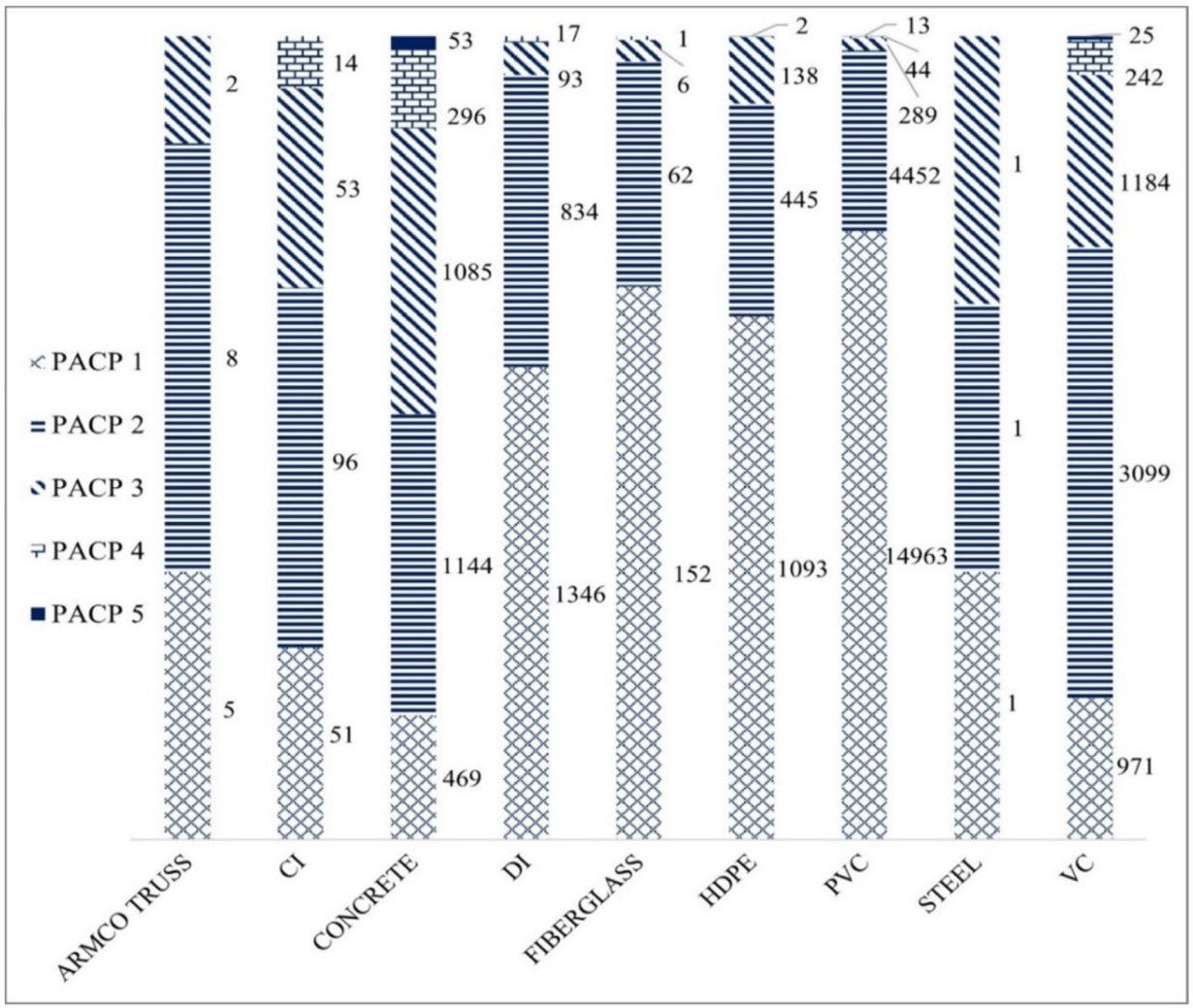

Fig. 4 Distribution of pipe materials and PACP scores

Table 4 Sample of processed data used to train the algorithms

Sl. No	Age	MAPSCO- GRID	Length	SUBAREA	Slope	Material	Diameter	PACP
1	13.5	46 H	426	MC04_04	0.018	PVC	8	1
2	14.4	119G	112	VC11_03	0.0057	PVC	8	1
3	15.2	106 S	396	VC11_01	0.0038	PVC	24	2
4	15.3	106U	444	VC11_01	0.0067	PVC	24	1
5	16.4	46 L	415	MC04_04	0.004	PVC	8	1
6	22.3	93G	844	VC09_01	0.0005	Concrete	54	2
7	33.3	103 H	95	SC09_05	0.0168	VCP	6	3
8	47.9	89 F	503	CF05_03	0.0301	Concrete	6	2
9	48	89 F	259	CF05_03	0.0206	VCP	6	3
10	67.4	47Y	203	MC03_06	0.0337	VCP	6	2

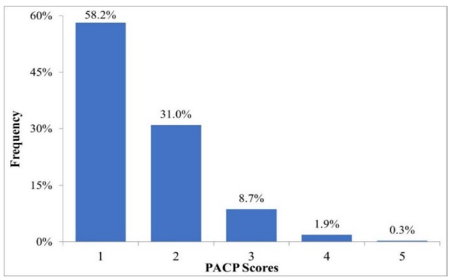

Fig. 5 Distribution of PACP scores

data to a logistic curve, which is then utilised to predict the likelihood of an event.

LR can be utilised when the outcome variable is categorical and has more than two categories. The term used to describe this sort of logistic regression is multinomial logistic regression. When using logistic regression for multi-class classification, the probability of one class is estimated in comparison to all other classes. The final model for a multinomial logistic regression can be represented by Eq. 1 (Agresti [1]).

$$\text{logit}\left[\frac{\pi}{1-\pi}\right] = \log\left(\frac{P(Y=1 \mid X_1, X_2, .X_p}{1- P(Y=1 \mid X_1, X_2, . X_p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_p x_p \quad (1)$$

where:
X1, X2, …, Xp are independent variables.
α is the intercept for ith category.
β is the regression coefficient.
k-Nearest Neighbor (k-NN).

The k-NN algorithm is a guided method commonly used for classification issues. The training dataset is the sole need for constructing a k-NN model. The algorithm identifies the nearest neighbours within the training dataset to categorise a new data point (Guo et al. [9]). The k-NN algorithm, in its simplest form, examines simply the nearest neighbour, The established result for this data point is subsequently utilised to formulate the forecast. Nevertheless, in order to enhance precision, it is possible to take into account an indefinite number of neighbouring data points, denoted as k (Müller and Guido [19]). Here, the k represents the number of neighbours, is set to 3.

Random forests

RF are widely used tree-based machine learning models known for their exceptional performance in handling big datasets (Loh [15]). RF is a machine learning technique that utilises ensemble learning. Ensemble learning involves combining multiple classifiers to address complex problems and improve the performance of the model. RF, or Random Forest, can be described as a compilation of diverse decision trees (DT), with each tree exhibiting slight variations from the others.

RF is known for its robustness in handling outliers and parameter spaces with large dimesionaity compared to other machine learning algorithms (Caruana and Niculescu-Mizil [5]). As a result, it is less prone to overfitting The Gini index (Gi) quantifies the ability of variables to predict outcomes in classification tasks (Alessia et al. [2]). The Gini index of a node 'n' is computed for a basic binary classification using Eq. 2.

$$Gi\,(n) = 1 - \sum {}_{j=1}^{2} \left(p_j\right)^2 \quad (2)$$

Here Pj is the relative frequency of clas j in the node n.

Evaluation metrics

In order to assess the accuracy of trained models in forecasting the state of sewer lines, it is necessary to validate and evaluate them. The literature study revealed that the bulk of the research employed commonly used evaluation measures, such as ROC curves and AUC values, for validation purposes. This work employs evaluation criteria, including accuracy, recall, precision, and F1-score, to evaluate the functionality of the trained models.

Cross validation

This is a frequently utilised validation method in any predictive problem. The core principle that underlies

Table 5 Python libraries used in this paper

Library Name	Description
Pandas	Access and alter numerical tables stored in spreadsheet files.
Scikit learn	This collection offers a diverse range of categorization algorithms.
Matplotlib	This library is often used for plotting graphs.
Seaborn	It is a library used for visualising data.

cross-validation is to withhold a piece of the input data during training of a model, and thereafter utilise this withheld fraction for testing the created model. The primary rationale of this metric is to mitigate overfitting and ensure that all classes are adequately represented during model training. In this study, a 5-fold cross-validation technique is employed, wherein the entire dataset is divided into 5 equal halves. Out of the total of 5 parts, 4 parts were allocated for training the model, while 1 part was reserved for testing the trained model.

Confusion matrix

This assessment metric is commonly accepted for assessing the functionality of a trained model. A confusion matrix is cross table that records the number of occurrences between two raters, the true/actual classification, and the anticipated classification. The confusion matrix arranges the accurately identified objects along the main diagonal, which extends from the top left to the bottom right (Grandini et al. [8]; Hossin and Sulaiman [13]). A confusion matrix provides a comprehensive assessment of a model's performance through visual examination.

Receiver operating characteristics

Figure 6 depicts a commonly used metric, known as a graph, that illustrates the performance of a classification model. This graph provides insights into the efficiency of the model. The graph displays the false positive rate (FPR) on the x-axis and the true positive rate (TPR) on the y-axis. TPR, is calculated by dividing the number of True Positives (TP) by the sum of True Positives and False Negatives (FN). On the other hand, FPR is calculated by dividing the number of False Positives (FP) by the sum of True Negatives (TN) and False Positives.

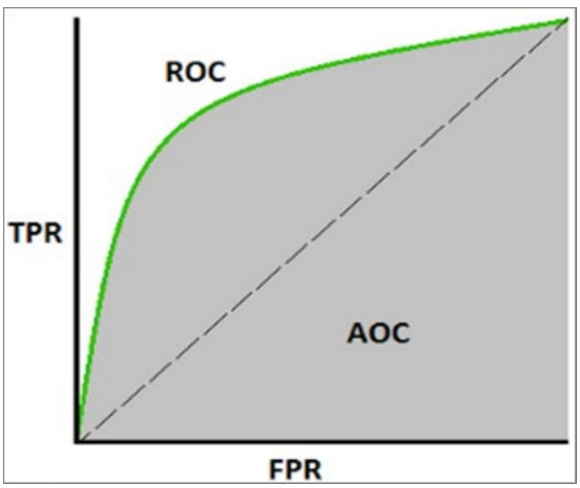

Fig. 6 ROC curve for a binary classification

ROC curve's area. Another intriguing statistic derived from the ROC curve is the area under the curve (AUC), represented by the shaded region in Fig. 6. The AUC value, which runs from 0 to 1, indicates the performance of a model in prediction. A higher AUC number indicates greater model performance.

Precision

The term "precision" refers to the proportion of correctly identified positive elements out of the total number of units that were predicted as positive, expressed as TP / (TP + FP). The precision of a model is critical when accurate predictions are necessary, especially when one class of the output variable is significantly less common than the other class (Grandini et al. [8]). Therefore, precision would be an important evaluation parameter to consider when selecting the model.

Recall

The fraction of positive elements accurately identified is measured as the ratio of true positives (TP) to the sum of true positives and false negatives (TP + FN). Assessing the model's capacity to accurately represent all positive aspects in the dataset is crucial (Grandini et al. [8]). The importance of precision and recall is evident, leading to the introduction of a new metric known as the F1-score.

However, this article requires the prediction of the condition of a pipe from a set of 5 distinct classes. Consequently, several measures of accuracy and completeness were determined for each category based on the confusion matrix for numerous classes, and the related F1-scores were computed.

The F1-score

The approach involved calculating the harmonic average of recall and precision, as demonstrated in Eq. 3. The F1-score is determined by taking a weighted average of precision and memory, with both factors carrying equal importance. This makes it a useful tool for finding the best balance between precision and recall (Grandini et al. [8]). The F1-score is a metric that measures the performance of a model on a scale of 0–1. A number of 1 indicates good functionality, while a lower value indicates poorer performance.

$$F1 - Score = 2 * \left(\frac{Precision * Recall}{Precision + Recall} \right) \quad (3)$$

Performance of developed models

Processed final dataset was trained with ML algorithms such as RF, k-NN and LR, and their results were compared using various metrics to determine the best

functioning model. Confusion matrices were constructed for each of the three models. In these matrices, the rows reflect the real class elements, which are based on PACP scores (1–5). The columns, on the other hand, indicate the anticipated class elements.

LR results

Confusion matrix for the model trained using LR algorithm was found to be

$$\begin{bmatrix} 3575 & 177 & 12 & 0 & 0 \\ 1383 & 611 & 65 & 0 & 0 \\ 171 & 336 & 79 & 0 & 0 \\ 25 & 71 & 25 & 0 & 0 \\ 5 & 11 & 5 & 0 & 0 \end{bmatrix}$$

In the confusion matrix above, PACP scores of 4 and 5 represnted by columns 4 and 5, respectively, are all recorded as zeros. This indicate that the model was unable to capture any of the sewer pipes that belonged to PACP scores 4 and 5. The confusion matrix can be visualized as shown in Fig. 7. The histogram displays the frequencies of each class that were incorrectly identified by the model as different classes. The greater the rate of incorrect classification of a model, the lower the reliability of that model. For example, approximately 5,000 observations were anticipated to have a PACP score of 1, however, over 1,000 observations were actually classified as having a PACP score of 2. Similarly, a significant number of observations that were identified as having a PACP score of 2 were actually incorrectly classified from classes 1, 3, and 4.

Based on estimated TPR and FPR from confusion matrix, ROC curves were plotted, and corresponding AUC values were obtained for all classes. Despite the model not making any predictions for classes 4 and 5, the estimated AUC values for these two classes were 0.89 and 0.80, respectively. The evaluation metrics are presented in Table 6.

k-NN results

Processed dataset was trained with k-NN algorithm, and the resulting confusion matrix was

$$\begin{bmatrix} 3248 & 509 & 37 & 5 & 0 \\ 1094 & 832 & 106 & 4 & 0 \\ 149 & 249 & 145 & 9 & 0 \\ 30 & 64 & 40 & 8 & 0 \\ 8 & 10 & 3 & 1 & 0 \end{bmatrix}$$

Similar to the LR confusion matrix, PACP score 5 observations are not captured by the trained model. On the other hand, the model captured PACP score 4 pipes to an extent. Nevertheless, the frequency of occurrences in both false positives and false negatives surpasses that of true positives. Hence, it may be inferred that the model is more likely to make incorrect classifications. The Fig. 8 depicts the error rate in prediction.

In addition to evaluation metrics, AUC values were calculated for 5 PACP classes, as presented in Table 7. Since

Table 6 Evaluation metrics for LR.

PACP Score	AUC	Precision	Recall	F1-score
1	0.78	0.693	0.950	0.801
2	0.66	0.507	0.297	0.374
3	0.85	0.425	0.135	0.205
4	0.89	0.000	0.000	0.000
5	0.80	0.000	0.000	0.000

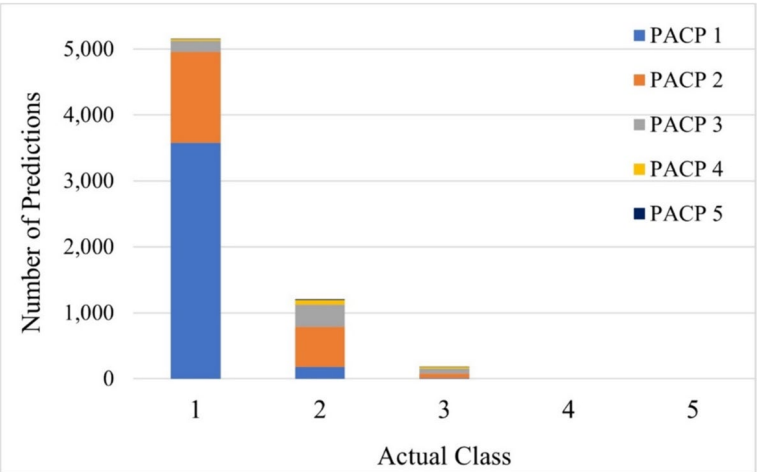

Fig. 7 Error prediction rate for LR model

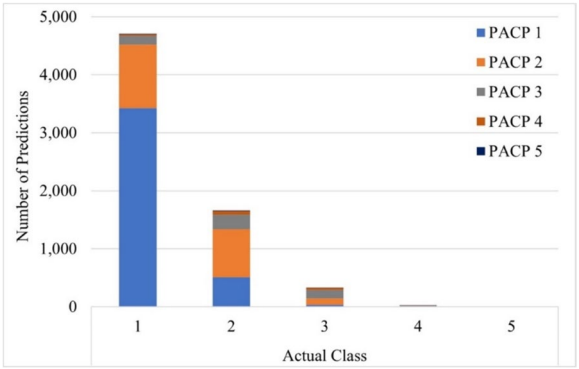

Fig. 8 Error prediction rate for k-NN model

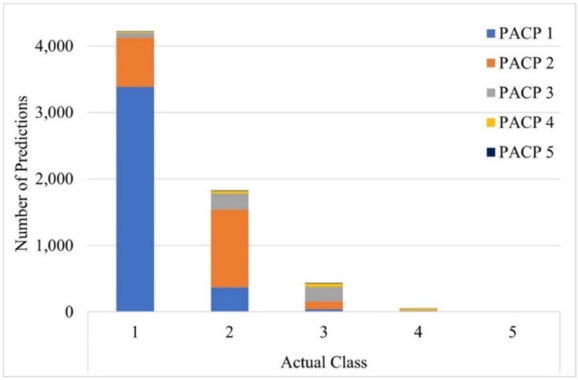

Fig. 9 Error prediction rate for RF model

Table 7 Evaluaton metrics for k-NN

PACP Score	AUC	Precision	Recall	F1-score
1	0.76	0.709	0.859	0.777
2	0.67	0.503	0.399	0.445
3	0.76	0.436	0.242	0.344
4	0.66	0.250	0.058	0.094
5	0.54	0.000	0.000	0.000

Table 8 Evaluation metrics for RF

PACP Score	AUC	Precision	Recall	F1-score
1	0.85	0.772	0.886	0.825
2	0.78	0.617	0.518	0.563
3	0.89	0.517	0.382	0.440
4	0.94	0.382	0.215	0.275
5	0.78	0.250	0.048	0.080

no observation was classified as PACP score 5, recall, precision, and F1-score are all zeros. It can be seen that majority of projected classifications under PACP score 4 were from other classes, which resulted a minimal recall score.

RF results

Thirdly, processed dataset was trained with RF algorithm and the resulted confusion matrix was

$$\begin{bmatrix} 3388 & 368 & 41 & 2 & 0 \\ 737 & 1176 & 117 & 6 & 0 \\ 75 & 236 & 222 & 18 & 1 \\ 20 & 40 & 53 & 27 & 2 \\ 5 & 8 & 5 & 3 & 1 \end{bmatrix}$$

As seen in the matrix, false positives and false negatives are comparatively lesser compared to confusion matrices of other two models. Interestingly, the model captured few observations under PACP score 5. However, it is evident that each class has misclassifications in it, which is illustrated in Fig. 9.

Based on estimated TPR and FPR from confusion matrix, ROC curves were plotted, and corresponding AUC values were obtained for all classes. Though recall, precision, and F1-score for PACP score 5 was minimal among 5 classes, AUC value was found to be 0.78, which is listed in Table 8. It should also be noted that AUC value of PACP class 4 was found to be the highest with 0.94 while the same class experienced major misclassification.

Though the RF model represented PACP class 5 unlike other models, total false negatives comparatively outnumbered true positives and hence, the recall was estimated to be 0.048. Due to merely 0 recall and a precision of 0.25, the resulted F1-score was nearly zero too, which indicated that the model is unreliable for PACP class 5.

Discussion and conclusion

The main contribution of this study is to employ multiclass classification and predict PACP scores of the pipes. Processed sanitary sewer dataset was tested with multiple supervised machine learning models and results were obtained. To effectively evaluate the functionality of generated models, precision, recall and F1-score values of all classes were averaged for each model. As shown in Fig. 10, LR model exhibited the lowest F1-score of 0.28 followed by k-NN and RF models with 0.33 and 0.44, respectively. Though it is evident from estimated metrics that LR model is not reliable for classification, the average AUC value was found to be 0.80, which is inconsistent. Therefore, it can be understood that AUC values cannot be considered as a single evaluation metric for condition prediction ML models.

It was revealed that the LR model was not able to capture both PACP 4 and 5 classes while k-NN was able to

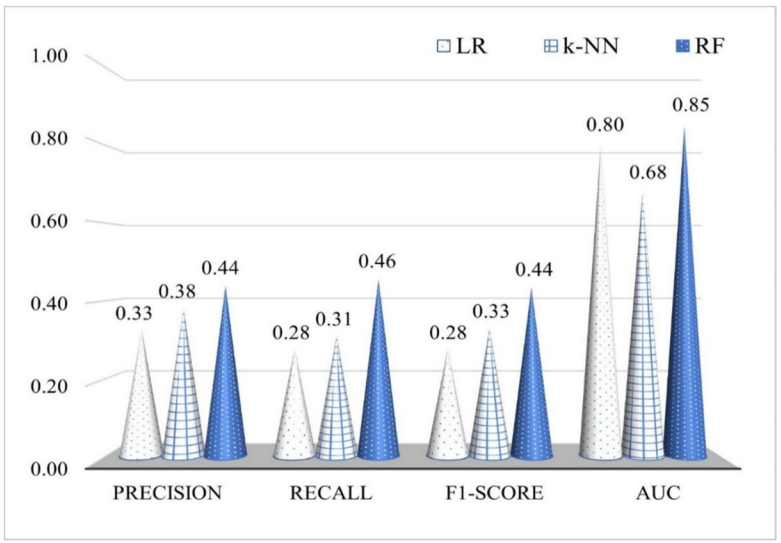

Fig. 10 Performance comparison between developed ML models

represent PACP class 4. The RF model was able to represent PACP class 5 but, false negatives outnumbered true positives. When compared to all three developed model, performance of RF model is found to be better than other two models. The developed models can be used by the asset managers during the inspections to assess the state of their sewage network and identify essential sewers that require immediate care. Implementing predictive models for assessing the condition of sanitary sewer pipes has significant practical implications for utility owners and city managers. These models enable a proactive maintenance approach, allowing for the early identification and prioritization of sewer pipe segments that require immediate attention. By focusing on pipes that are more likely to fail, utility companies can reduce the frequency and scope of extensive CCTV inspections, which are both costly and labor-intensive. This not only leads to significant cost savings but also optimizes the use of financial and human resources. Furthermore, predictive models enhance the overall asset management framework by providing a data-driven basis for decision-making, leading to more informed strategies for maintaining and upgrading sewer infrastructure. The poor performance of the ML models is due to skeweness of the data. For future studies, the research will consider the sampling techniques to remove the skeweness of the data.

Acknowledgements
Not applicable.

Authors' contributions
"Conceptualization, K.L., M.N., and S.K.; methodology, K.L; software, K.L.; validation, K.L., M.N., and S.K.; formal analysis, K.L.; writing—original draft preparation, K.L., M.N., S.K., P.K.M., and A.P; writing—review and editing, K.L., M.N., S.K., P.K.M., and A.P visualization, K.L; supervision, M.N., and S.K. All authors have read and agreed to the published version of the manuscript."All authors read and approved the final manuscript

Funding
This research receieved no external funding.

Availability of data and materials
The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare no competing interests.

Received: 2 January 2024 Revised: 9 July 2024 Accepted: 15 July 2024
Published online: 29 July 2024

References

1. Agresti A (2007) An introduction to categorical data analysis. Wiley, Hoboken, NJ
2. Alessia S, Antonio C, Aldo Q (2017) Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: a systematic review. Front Aging Neurosci 9:329
3. Ana EV, Bauwens W (2010) Modeling the structural deterioration of urban drainage pipes: the state-ofthe-art in statistical methods. Urban Water J 7(1):47–59
4. Ariaratnam ST, El-Assaly A, Yang Y (2001) Assessment of Infrastrucure Inspection needs using logistic models. J Infrastruct Syst 7(4):160–165
5. Caruana R, Niculescu-Mizil A (2006) An Empirical Comparison of Supervised Learning Algorithms. Proceedings of the 23rd International Conference on Machine Learning, (pp. 161–168). Pittsburgh
6. EPA (2002) Asset Management for Sewer Collection systems. Office of Wastewater Management, Washington, DC
7. EPA (2015) *Condition Assessment of Underground Pipes* With excerpts from: Condition Assessment of Wastewater Collection Systems,

EPA/600/R-09/049. Retrieved from EPA: https://www3.epa.gov/region1/sso/pdfs/condition-assessment-underground-pipes.pdf
8. Grandini M, Bagli E, Visani G (2020) Metrics for Multi-class classification: an overview. BO), Italy: Università degli Studi di Bologna, Bologna
9. Guo G, Wang H, Bell D, Bi Y, Greer K (2003) k-NN Model-Based Approach in classification. Move Meaningful Internet Syst, 986–996
10. Hahn M, Palmer R, Merrill S, Lukas A (2002) Expert System for prioritizing the inspection of sewers: knowledge base formulation and evaluation. J Water Resour Plan Manag 128(2):121–129
11. Harvey RR, McBean EA (2014) Comparing the utility of decision trees and support vector machines when planning inspections of linear sewer infrastructure. J Hydroinformatics 16(6):1265–1279
12. Hernandez N, Caradot N, Sonnenberg H, Rouault P, Torres A (2017) Support Tools to Predict the Critical Structural Condition of Uninspected Sewer Pipes in Bogota D.C. *The Leading Edge Sustainable Asset Management of Water and Wastewater Infrastructure Conference* Trondheim, Norway
13. Hossin M, Sulaiman M (2015) A Review on Evaluation Metrics for Data Classification Evaluations. Int J Data Min Knowl Manag 5(2):1
14. Loganathan K (2021) Development of a Model to Prioritize Inspection and Condition Assessment of Gravity Sanitary Sewer Systems. Dissertation, University of Texas at Arlington, Arlington, TX
15. Loh W-Y (2014) Fifty years of classification and regression trees. Int Stat Rev 82(3):329–348
16. Malek Mohammadi M, Najafi M, Kaushal V, Serajiantehrani R, Salehabadi N, Ashoori T (2019) Sewer Pipes Condition Prediction Models: A State-of-the-Art Review. Infrastructures 4(4):64
17. Malek Mohammadi M, Najafi m, Tabesh A, Riley J, Gruber J (2019b) Condition Prediction of Sanitary Sewer pipes. ASCE Pipelines, 117–126
18. Mohri M, Rostamizadeh A, Talwalkar A (2018) *Foundations of Machine Learning (second edition)* Cambridge. The MIT Press, MA
19. Müller AC, Guido S (2016) Introduction to machine learning with Python: a guide for data scientists. O'Reilly, Boston
20. Najafi M, Gokhale S (2005) Trenchless Technology. McGraw-Hill, New York
21. Najafi M, Gokhale S (2021) *Trenchless Technology: Pipeline and Utility Design, Construction, and Renewal*, 2nd edn. McGraw-Hill, New York
22. Najafi M, Kulandaivel G (2005) Pipeline Condition Assessment Prediction using neural network models. ASCE Pipelines, 767–781
23. NASSCO. (2018), January Pipeline Assessment Certificate Program
24. Shmilovici A (2023) Support Vector machines. In: Rokach L, Maimon O, Shmueli E (eds) Machine learning for Data Science Handbook. Springer, Cham
25. Syachrani S, Jeong HS, Chung CS (2013) Decision tree–based deterioration model for buried Wastewater Pipelines. ASCE J Perform Constructed Facilities 27(5):633–645
26. Tscheikner-Gratl F, Caradot N, Cherqui F, Leitão JP, Ahmadi M, Langeveld JG, Clemens F (2020) Sewer asset management – state of the art and research needs. Urban Water J 16(9):662–675
27. Vladeanu GJ (2018) Wastewater Pipe Condition and Deterioration modeling for risk-based decision making. Louisiana Tech University, Louisiana
28. Wirahadikusumah R, Abraham D, Iseley T (2001) Challenging issues in modeling deterioration of combined sewers. ASCE J Infrastructure Syst 7(2):77–84
29. Wright LT, Heaney JP, Dent S (2006) Prioritizing Sanitary sewers for Rehabilitation using least-cost classifiers. Asce J Infrastructure Syst 12(3):174–183

Publisher's Note