Journal of Civil Engineering and Architecture 19 (2025) 105-119
doi: 10.17265/1934-7359/2025.03.001

DAVID PUBLISHING

Development of Machine Learning Based Prediction Models to Prioritize the Sewer Inspections

Madhuri Arjun[1] and Arjun Nanjundappa[2]

1. Department of Civil Engineering, University of Texas at Arlington, 1221 West Mitchell Street, Room # 313, Arlington, TX 76013, USA

2. Mavenir Systems, 1700 International Pkwy, Richardson, TX 75082, USA

Abstract: Sewer pipe condition assessment by performing regular inspections is crucial for ensuring the systems' effective operation and maintenance. CCTV (closed-circuit television) is widely employed in North America to examine the internal conditions of sewage pipes. Due to the extensive inventory of pipes and associated costs, it is not practical for municipalities to conduct inspections on each sanitary sewage pipe section. According to the ASCE (American Society of Civil Engineers) infrastructure report published in 2021, combined investment needs for water and wastewater systems are estimated to be $150 billion during 2016-2025. Therefore, new solutions are needed to fill the trillion-dollar investment gap to improve the existing water and wastewater infrastructure for the coming years. ML (machine learning) based prediction model development is an effective method for predicting the condition of sewer pipes. In this research, sewer pipe inspection data from several municipalities are collected, which include variables such as pipe material, age, diameter, length, soil type, slope of construction, and PACP (Pipeline Assessment Certification Program) score. These sewer pipe data exhibit a severe imbalance in pipes' PACP scores, which is considered the target variable in the development of models. Due to this imbalanced dataset, the performance of the sewer prediction model is poor. This paper, therefore, aims to employ oversampling and hyperparameter tuning techniques to treat the imbalanced data and improve the model's performance significantly. Utility owners and municipal asset managers can utilize the developed models to make more informed decisions on future inspections of sewer pipelines.

Key words: Sanitary sewers, asset management, pipe inspection, ML algorithms, condition prediction models.

1. Introduction

The U.S. underground sewer systems are a significant part of municipal infrastructure, comprising thousands of miles of pipelines designed to carry and transport domestic sewage and stormwater runoff to the treatment plants [1]. Most sewer pipes in the U.S. operate using a gravity-based system. Gravity sewer systems rely on the land's natural slope to transport wastewater from higher to lower elevations, directing it to a treatment facility or a disposal point. The United States has over 800,000 miles of publicly owned sewer pipelines and over 500,000 miles of privately owned sewer laterals; 240 million Americans have access to 14,748 wastewater treatment facilities. It is anticipated that 56 million additional people will use concentrated treatment facilities by 2032 [2].

According to EPA (Environmental Protection Agency) [3] research, up to half of the buried assets in investigated systems may be past the midpoint of their service lives. Most municipal sewers are a significant portion of the wastewater infrastructure in the United States and are over a century old; aging, chemical, and environmental variables all result in at least 23,000 to 75,000 sanitary sewer overflows annually [4].

On the most recent infrastructure report card, published by the ASCE (American Society of Civil Engineers) [5] in 2021, the wastewater infrastructure received a "D plus" grade. According to ASCE, water and wastewater systems in the United States are readily

Corresponding author: Madhuri Arjun, PhD, assistant professor, research fields: pipeline engineering, construction management.

Arjun Nanjundappa, MS, Director, research fields: system engineering, AI – ML programming.

aging, and an investment deficit of $150 billion must be addressed by 2025 to keep up with the needs [5]. Besides, the population of the U.S. is expanding and changing geographically. This demands investment in new infrastructure and maintenance of existing infrastructure in places with declining populations and confined budgets [5].

According to AWWA (American Water Works Association) [6], various municipalities and agencies prioritize sewage restoration rather than adding new sewer lines to meet growth or upgrading treatment plants. Inadequate maintenance and poor asset management methods raise the danger of inflow and infiltration, sanitary sewage overflows, and sinkholes. Failure to manage clean sewer systems could harm human health while leading to expensive property damage and emergency repairs [7].

In contrast to reliant maintenance procedures used by certain municipalities after pipe breakdown, preventive maintenance should involve inspection and maintenance activities before failure or irreversible deterioration [8].

- Structural—cracks, fractures, breaks, and so on;
- Hydrostatic—flooding, encrustation, and grease;
- Corrosion—chemical and external corrosion;
- Erosion;
- Operational problems—roots, blockages, debris, and so on.

As previously stated, sewer pipes are a vital component of wastewater systems since they connect points of wastewater generation to treatment plants. There may be a decline in structural and operational efficiency as sewer systems age. Old or deteriorated pipes can cause problems for people's health, the environment, and the economy [9].

Maintenance and rehabilitation methods are essential in sustaining the pipeline's operation at an acceptable level of service and offering cost-effective ways to avoid unplanned failures. Previously, sewer pipe repair or rehabilitation was only done when a pipe collapsed or failed. The current trend, however, is to repair and manage pipe systems before they collapse. Municipalities and utilities have begun to adopt asset management systems to attain this goal. Infrastructure asset management is a thorough and efficient process [10]. An efficient asset management strategy can include several techniques to assist utility owners and municipalities understand the time and related costs of deteriorating pipe repair, rehabilitation, or replacement [11].

Sewer pipeline deterioration depends on many factors and steps, making it harder for municipalities to locate collapse-prone pipes. In recent years, sewage pipeline inspection and monitoring have intensified to prevent further collapse and failure. Hence, pipe deterioration models that predict sewer pipeline conditions must be developed [12]. This dissertation examines statistical and AI (artificial intelligence) algorithms for sanitary sewage pipe condition prediction. Clean sewer pipe effect aspects will also be discussed.

The primary objectives of this research work include:

- To identify critical variables affecting sanitary sewer pipeline conditions.
- To develop an AI-based prediction model capable of forecasting the deterioration of sewer pipes.

The scope of this research is restricted to the study of sanitary sewers with gravity flow using the PACP (Pipeline Assessment Certification Program) scores that the operators record by carrying out the CCTV (closed-circuit television) inspection for modeling the deterioration of pipe systems. The condition of sewer pipes is categorized based on the PACP developed by the NASSCO (National Association of Sewer Service Companies). Table 1 presents the scope of this research.

PACP was developed in 2001 by the NASSCO in association with WRc (Water Resources Center) to design a standard for sewer condition assessment. PACP aims to construct a database to accurately identify, plan, prioritize, manage, and renovate sewer pipe assets based on condition assessment.

Table 1 Scope of the research.

Included	Not included
Sanitary sewer pipes	Stormwater pipes
Gravity sewer pipes	Force main sewer pipes
Inspected pipes based on PACP guidelines	Inspected pipes based on other guidelines
VCP, PVC, RCP, UnReinCONC, RPM, DI, CI, AC, HDPE, PCCP, FRP, CLC, CMP	Other not included sewer pipes
Sanitary sewer pipes without any repair or rehabilitation history	Pipe segments that have a history of pipe maintenance

Table 2 PACP condition rating [13].

PACP	Description	Predicted failure time
1	Excellent	Failure unlikely soon
2	Good	20 years or more
3	Fair	10-20 years
4	Poor	5-10 years
5	Immediate attention	Failed or imminent failure within the next five years

According to the NASSCO coding system, pipe defects and features can be classified into five categories. The defect classification includes classes for (1) continuous defects, (2) structural defects, (3) operational and maintenance, (4) construction features, and (5) other features [13].

Several factors, such as the significance of the defect, the extent of the damage, and the percentage of restriction to flow capacity or wall loss due to deterioration, are used to assign grades. The final condition rating is derived from the categories of structural, O&M (operation, and maintenance). Table 2 presents the PACP condition rating representing the NASSCO 2015 manual.

2. Literature Review

Several eminent researchers in the United States have developed condition prediction models to identify the critical factors that influence the deterioration of sanitary sewer pipelines. The developed deterioration models utilize statistical methods and AI-based algorithms. However, a single standard model has yet to be created by collecting data from different geographical areas in the United States. This proves that city municipalities cannot employ the prediction models developed by past researchers to prioritize inspection operations on sewer pipes [14]. One of the most critical limitations of current sewer prediction models has been the need for data from different geographical locations to train and validate reliable models. Several contributors suggested improving sewer pipe condition prediction models from various perspectives, which includes below factors:

- Najafi and Kulandaivel [15] said that the neural network model for sewer pipe deterioration could be improved by adding more historical input variables, such as surface load, groundwater, bedding conditions, soil corrosion, stability, and sewer location.
- Chughtai [16] recommended incorporating more variables like soil type and its conditions, predicting the sewer pipe deterioration models. Future research should investigate the application of further prediction models.
- Mashford et al. [17] suggested that pipe length and depth data must be incorporated as independent variables in developing prediction models.
- Sousa et al. [18] proposed using higher-level deterioration models and comparing the results to ML (machine learning) and neural network models.
- Kabir et al. [19] presented that the developed sewer structural condition prediction models may be improved by assessing the effects of additional independent variables, such as sewer function, groundwater level, soil type, road class, and initial quality of construction.

Table 3 Summary of developed prediction models [20].

	Authors	Year	Model	Variables included	Condition assessment standard	Condition rating output	Number of data
1	Ariaratnam et al. [21]	2001	• Logistic regression	Age, Material, Diameter, Depth	1,2,3,4,5	WRc	748
2	Najafi and Kulandaivel [15]	2005	• Neural network	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	-
2	Tran et al. [22]	2006	• Neural network	Age, Diameter, Depth, Length, Slope	1,2,3	WSAA	583
3	Tran et al. [23]	2007	• Neural network • Multiple discrimination analysis	Age, Diameter, Depth, Length, Slope	1,2,3	WSAA	150
4	Chughtai and Zayed [24]	2008	• Linear regression	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	WRc	-
5	Tran et al. [25]	2009	• Neural network • Ordered probit model	Age, Diameter, Depth, Length, Slope	1,2,3	WSAA	417
6	Khan et al. [26]	2010	• Neural network	Age, Diameter, Depth, Length, Slope	1,2,3,4,5	WRc	200
7	Lubini and Fuamba [27]	2011	• Logistic regression	Age, Diameter, Depth, Length, Slope	1,2,3	PACP	459
8	Mashford et al. [28]	2011	• Support vector machine (SVM)	Age, Material, Diameter, Slope	1,2,3	PACP	1,441
9	Salman and Salem [29]	2012	• Ordinal regression • Logistic regression • Binary regression	Age, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	11,373
10	Syachrani et al. [30]	2013	• Decision tree • Neural Network	Age, Material, Diameter, Length, Slope	1,2,3,4,5	PACP	52,855
11	Sousa et al. [18]	2014	• Neural network • Support vector machine • Logistic regression	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	745
12	Harvey and McBean [31]	2014	• Random forest • Decision Tree • Support vector machine	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	WRc	1,825
13	Bakry et al. [32]	2016	• Multiple regression	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	84
14	Gedam et al. [33]	2016	• Linear regression	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	155
15	Hernandez et al. [34]	2017	• Logistic regression • Random forest	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	4,327
16	Kabir et al. [19]	2018	• Bayesian logistic regression	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	12,728
17	Laakso et al. [35]	2018	• Binary logistic regression • Random forest	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4	EN-13508-2	6,700
18	Malek Mohammadi [36]	2019	• Logistic regression • k-NN • XGBoost	Age, Material, Diameter, Length	1,2,3,4,5	PACP	20,282
19	Mazumder et al. [37]	2020	• Decision tree • Random forest • AdaBoost, XGBoost, LGBoost, CATBoost	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	959
20	Loganathan [38]	2021	• Logistic regression • k-NN • Random forest	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	32,751
21	Shirkhanloo [39]	2022	• k-NN • Random forest • AdaBoost, XGBoost • Gradient Boost	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	3717

- Malek Mohammadi [36] stated that a prediction model must be able to predict each of the five condition levels independently instead of transforming them into binary classes.
- Loganathan [38] developed the prediction models by adopting advanced ML algorithms using data collected from one city municipality. The research suggested that models must be validated with the inspection data collected from different cities.
- Shirkhanloo [14] used supervised learning algorithms to develop prediction models by gathering data from a single municipality. The researcher pointed out that the inspection data from more cities with variables like pipe diameter, material, length, depth, soil, and PACP ratings are required to develop the prediction models that the city municipalities can employ to prioritize the sewer pipe inspection.

From the studies mentioned earlier and the developed models, it is shown that there is a significant knowledge gap in identifying the vital variables affecting the deterioration of sewer lines. It was also presented that most studies needed more variations in the data collected from city municipalities and were found to be restricted. This is a significant limitation of the prediction models developed above as they are based on the single city analysis. The summary of the developed prediction models to date can be seen as shown in Table 3.

3. Methodology

The AI models developed in this paper are used to predict the condition rating of individual sewer lines by considering the physical features of the pipelines and various environmental factors. These significant variables may lead to the eventual deterioration of sewer pipelines. As illustrated in Fig. 1, the following steps are followed in this methodology to achieve the intended result of the research.

The study implements supervised learning techniques to construct a prediction model.

3.1 Model Building

Regression and classification are the main supervised ML methods. Regression occurs when many independent factors must predict a continuous dependent variable [39]. The dependent or outcome variable in this study is categorical with five classifications. This study develops a model using classification ML techniques. Models are trained using processed data, as discussed in the previous chapter. Python, a prominent data science programming language, is utilized to create prediction models in this study. Python's open-source nature and many free add-on packages make it appealing.

The number and type of independent variables, as well as the dependent variable, affect the sewer conduit prediction models. It is crucial to select a predictive model designed for predicting dependent factors with multiple classes, given that the dependent variable in this study is the condition rating of sewer pipelines, which has been organized into five categories. Therefore, the best models for this research were selected based on their ability to predict multi-categorical dependent variables.

Using Python software, this dissertation develops a statistical model utilizing logistic regression and classification. For analyzing datasets with two or more discrete outcome variables, the classification model is the most widely used. Decision tree and random forest models with default parameters were the first set of models developed in this research to evaluate their performance score (accuracy and F1 score).

The second set of models devised in this study are hyperparameter tuning of decision trees and random forest algorithm to improve the performance. Python was chosen for this study because it is an open-source programming language with various free add-on libraries. As a third type of model, tree-based boosting models are created. They are among the most effective learning techniques presented and are intended for classification problems. In this dissertation, decision tree, bagging techniques, such as random forest, and boosting techniques, such as AdaBoost, Gradient Boosting Tree, and XGBoost, are developed and explained.

3.2 Model Selection

Model selection is crucial to statistical analysis since numerous factors affect regression models. Sewer pipe deterioration models depend on information data, independent variables, and dependent variables. As mentioned, sewer pipe values classify condition prediction scales. Choose a predictive model that can forecast categorical dependent variables.

This research study aims to forecast the future condition states of individual sewer pipelines. It is seen that the condition states of sewer pipes are typically described as discrete or categorical values which are not serial numbers and are classified with 5 different classes; consequently, linear and exponential regressions are not appropriate for predicting categorical variables because they minimize the total number of squared distances between the predicted and actual condition ratings [40]. So, the classification type of regression techniques is employed to construct the model.

In this research, the most ideal models are selected on the model performance in predicting categorical dependent variables, model's capability to be trained by nominal variables and model's end results.

3.3 Imbalanced Data Correction by Oversampling Method

The data collected from the sewer inspection have pipe PACP score of 4 and 5 in its rare occurrence when compared to the PACP scores of 1, 2 and 3. This variance in the dataset is termed an imbalanced state. When imbalanced data are used in traditional classification algorithms, it can often lead to poor model performance [41]. In general, the minority class must be accorded with a higher priority when dealing with unbalanced datasets, as the repercussions of misclassifying a minority class are exponentially more significant than those of the other classes. In this study, PACP scores of 4 and 5 are given more weight because interpreting it as scores of 1, 2 and 3 would be more detrimental.

Data scientists and researchers explored different techniques to treat this imbalance in dataset and found that classification algorithms like logistic regression, support vector machine, and decision tree can be successfully used in training the dataset [42]. This study employs the data resampling method to treat the imbalance datasets which is the most effective in replicating or removing the data points to make the majority class and minority class meet the requirement. The data resampling technique is divided into (1) random under-sampling and (2) random over-sampling.

This paper explores on finding optimal hyperparameters for a model, known as hyperparameter tuned. Hyperparameters are the parameters that govern the entire training process. The hyperparameter values are set at the beginning of the learning process. Selecting optimal hyperparameters can lead to increases in the overall model's performance and can help in reducing both overfitting and underfitting and will have a substantial effect on the model's performance.

Finding the optimal set of hyperparameter values for models with many hyperparameters can be a time-consuming endeavor. To make the procedure more efficient, two of the most prevalent methods are available in sklearn: Grid-Search and Random-Search.

In this research, five classification decision tree models are developed to create the prediction models namely—Decision Tree, Random Forest, AdaBoost, Gradient Boost, XGBoost algorithms with imbalanced and oversampled datasets by employing default parameters and hyperparameter tuning technique. Table 4 presents the built-in hyperparameters in python for tree-based models developed in this study. While building the tree-based models, the below mentioned hyperparameters are set to get better performance.

Table 4 Built-in hyperparameters for tree-based models.

Decision trees	Random forest	Adaboost	Gradient boost	XGBoost
max_depth	n_estimators	base_estimator	Learning_rate	Learning_rate
min_samples_split	max_features	n_estimators	gamma	gamma
min_samples_leaf	max_depth	learning_rate	scale_pos_weight	scale_pos_weight
max_features	min_samples_split	algorithm	N_estimators	cosample_bytree
class-weight	min_samples_leaf	classes	max_depth	colsample_bylevel
	bootstrap	estimator_weights	min_samples_split	colsample_bynode
		estimator_errors	max_features	max_features

3.4 Evaluation Matrics

To determine the effectiveness of trained models in predicting the condition of sewer pipelines, it is essential to validate and assess their performance. A review of existing literature indicates that most prior studies have relied on widely used evaluation metrics, such as ROC (Receiver Operating Characteristic) curves and AUC (Area under the Curve) values, for model validation. In this study, additional performance measures—including accuracy, recall, precision, and F1-score—are utilized to comprehensively assess the efficiency of the trained models.

3.5 Cross Validation

Cross-validation is a widely used technique for validating predictive models. The fundamental concept behind this method is to set aside a portion of the dataset during the training phase and later use it to evaluate the model's performance. This approach helps prevent overfitting and ensures that all classes are fairly represented throughout the training process. In this study, a 5-fold cross-validation method is applied, where the dataset is divided into five equal segments. Among these, four parts are used for training, while the remaining one is utilized for testing.

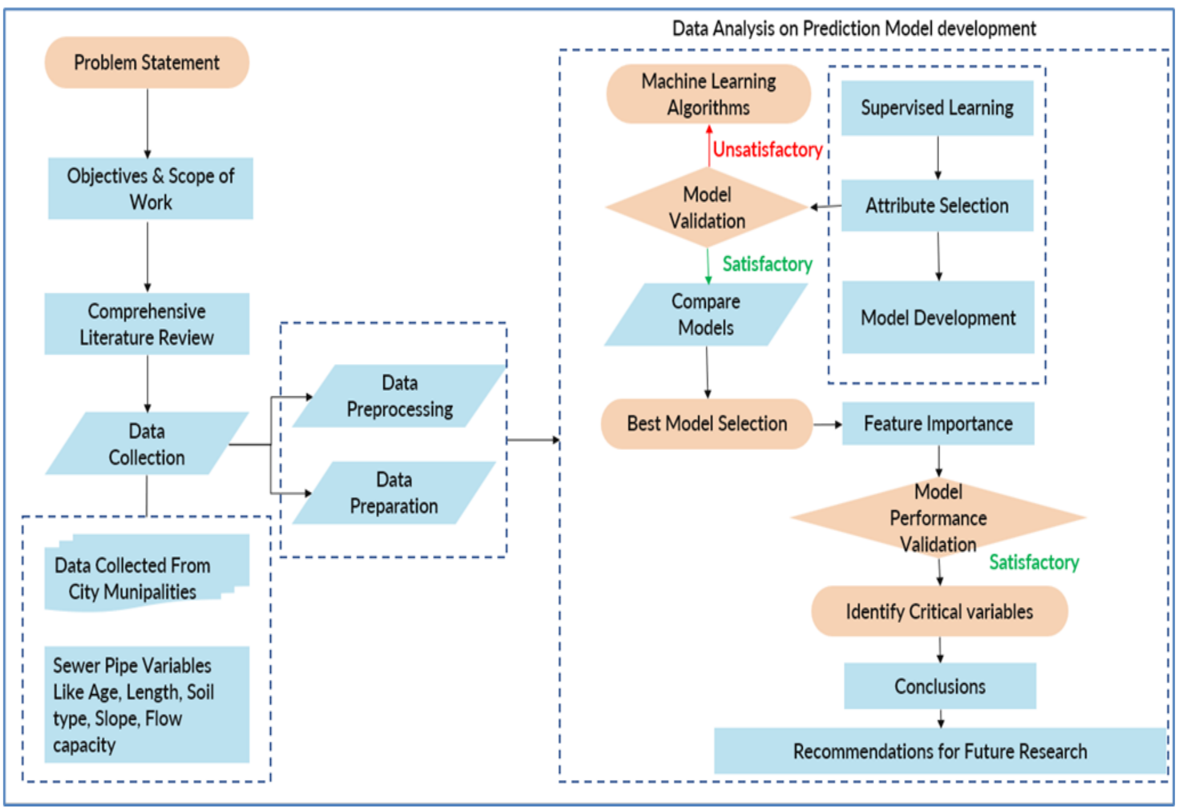


Fig. 1 Research methodology [20].

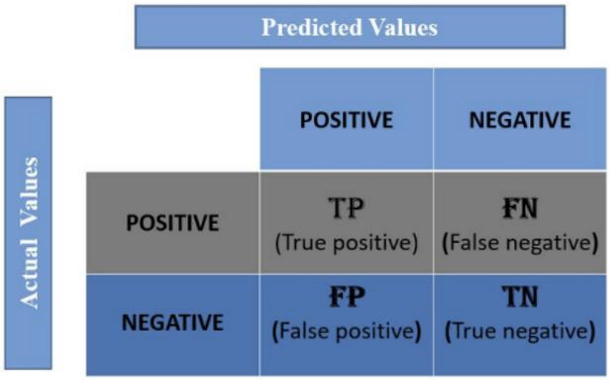


Fig. 2 Confusion matrix for binary classification [36].

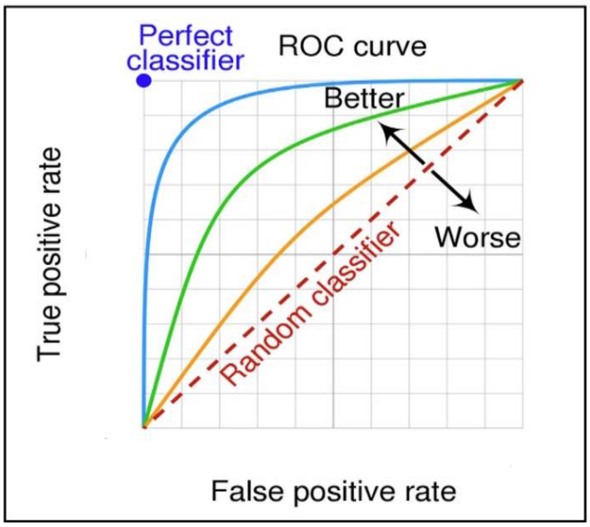


Fig. 3 ROC curve for binary classification [36].

3.6 Confusion Matrix

The confusion matrix is a widely recognized tool for evaluating the effectiveness of a trained model. It is a cross-tabulation that captures the frequency of occurrences between the actual classifications and the predicted classifications. The correctly identified instances are positioned along the main diagonal, running from the top left to the bottom right of the matrix [43]. By offering a visual representation of model performance, the confusion matrix enables a detailed assessment of how well the model distinguishes between different classes. Fig. 2 presents the confusion matrix for binary classification.

3.7 ROC Curve

ROC is a widely used evaluation metric, commonly represented as a graph, which illustrates the effectiveness of a classification model. This visualization offers valuable insights into the model's predictive capability.

The ROC curve plots the FPR (false positive rate) on the *x*-axis against the TPR (true positive rate) on the *y*-axis. The TPR is determined by dividing the number of TPs (true positives) by the total sum of TPs and FN (false negatives). Conversely, the FPR is calculated as the ratio of FPs (false positives) to the total number of TN (true negatives) and FPs [44]. Fig. 3 shows the ROC curve for binary classification.

Table 5 shows the evaluation metrics used to calculate the precision, recall, accuracy and F1 score.

Precision and recall are essential metrics used to assess a model's effectiveness in identifying positive instances. Precision measures the proportion of correctly predicted positive cases; out of all instances the model is classified as positive. On the other hand, recall represents the percentage of actual positive cases that the model successfully identified.

Accuracy refers to the proportion of correctly predicted instances out of the total predictions made by the model. It is commonly used as a key performance metric in classification tasks to evaluate overall model effectiveness. F1 score is derived by taking a balanced average of precision and recall, where both metrics are given equal significance. It evaluates a model's performance on a scale ranging from 0 to 1, with a value closer to 1 indicating strong predictive accuracy, while lower values suggest weaker performance.

4. Constructing a Prediction Model

Prediction of sanitary sewer pipes' condition is a well-established concept. With the development of ML and sophisticated statistical analysis utilizing computers, algorithms, or artificial intelligence, researchers have conducted numerous studies to predict the condition of sewer pipes. The two main categories of ML are supervised and unsupervised. Most data analysis in numerous condition prediction-related studies is classified as supervised learning. This research uses the supervised

Table 5 Evaluation metrics used to calculate the precision, recall, accuracy and F1 score [20].

		Predicted values		
		True	False	
Actual	True	TP	FN Type 1 Error	$\text{Recall} = \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$
	False	FP Type 1 Error	TN	$\text{Specifcity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$
		$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$		$\text{Acuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$ $\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Table 6 Descriptive statistics of pipe variables [20].

	Pipe age	Pipe diameter	Pipe length	Pipe slope	PACP score
Count	4,268	4,268	4,268	4,268	4,268
Mean	40.65	11.26	269.00	0.66	1.96
Std.	22.39	9.97	223.00	1.18	1.33
Min	1.00	4.00	0.50	-0.02	1.00
25%	19.00	6.00	114.10	0.04	1.00
50%	45.00	8.00	221.00	0.28	1.00
75%	52.00	10.00	371.00	0.60	3.00
Max	121.00	96.00	2,054.00	14.12	5.00

learning method to develop the prediction model, and the target variable is evaluated against the most influential independent variable. Three supervised boosting algorithms, Gradient Boost, AdaBoost, and XGBoost, were employed in this research to develop sewer prediction models.

4.1 Data Description

Sewer pipe inspection data for 2022 are collected from the southcentral region of the United States. The sewer pipe database inventory includes pipe variables like age, diameter, materials, length, slope, and soil type, termed independent variables. The PACP score of the data is the target or dependent variable with a rating from 1 to 5. Table 6 shows the descriptive statistics of pipe variables considered in this study.

This study's EDA (exploratory data analysis) showed that the PACP scores of 1, 2, and 3 have more

frequency; in contrast, PACP scores of 4 and 5 have a rare frequency. We refer to this variation in the dataset as an imbalanced state. Poor model performance frequently results from using unbalanced data in classical classification techniques [2]. When working with imbalanced datasets, it is necessary to give the minority class a higher priority because the consequences of incorrectly classifying minority classes are exponentially more significant than those of the other classes. A PACP score of 4 or 5 is given more weight in this study since it would be more accurate to interpret it as a score of 2 or 3.

After investigating several methods to address this imbalance in the dataset, the most efficient technique for duplicating data points to ensure that the majority class and minority class satisfy the requirements is data resampling, used in this work to handle the imbalanced datasets.

This study used the SMOTE (synthetic minority over-sampling technique) approach to evaluate the oversampling methodology for dealing with an imbalanced dataset. Several ML techniques were used to generate prediction models, including AdaBoost, Gradient Boosting, and XGBoost with the default parameter, as well as AdaBoost, Gradient Boosting trees, and XGBoost with the oversampled hyperparameter tuned. This study used a confusion matrix, accuracy, and macro F1 score on imbalanced datasets to compare the resampling and hyper-tuning techniques. We utilized the imbalanced-learn package in Python to implement resampling techniques.

This research employed the adaptive boosting (Adaboost) classifier and gradient boost classifier from the scikit-learn package in Python as the classification algorithm to evaluate the results of each resampling technique. The logic underlying the selection of that algorithm is to firstly compare the performance of all the boosting algorithms in their default hyperparameter configuration. A randomized grid search was used to tune the classification algorithm, and 40 sets of hyperparameter values were fitted with 3-fold CV (cross-validation), totaling 120 model fits. The candidate hyperparameter values were as follows: the options of max tree depth (max depth) values were 100, 140, 180, 220, and none; the options of the number of trees (*N* estimator) were 250, 500, 750, 1,000, the options of the max feature, when considering the best node split was "auto", which takes all features at the split; and sqrt (i.e., square root), which takes the square root of the total number of features in a tree. After obtaining the optimal hyperparameter value, each classification model was trained, tested, and, validated with repeated stratified K-fold CV. A 10-fold CV was performed repeatedly 4 times, totaling 40 fits, accounting for the class proportion of the targeted variable. Then, the mean and SD (standard deviation) of the accuracy, precision, recall, and Macro F1 score were extracted from 40 model fits.

5. Results

Table 7 lists the optimized hyperparameter values tuning result of randomized grid search oversampling technique. Table 8 shows the classification models output of each boosting algorithm developed with default and hyperparameter tuning.

It can be seen that, AdaBoost classification model with default parameter tuning and randomized hyperparameter tuning showed the low accuracy and macro F1 score as 0.675 and 0.676 and Fig. 4 depicts the confusion matrix of the AdaBoost classification models for the test set. The values in all classes 1 to 5 are wrongly predicted to the actual values, hence resulting in low performance of the developed model. By utilizing the hyperparameter tuning the model by oversampling technique, the performance of the training set was improved but the model showed overfitting.

Table 7 Optimized hyperparameter values for each model from randomized grid search.

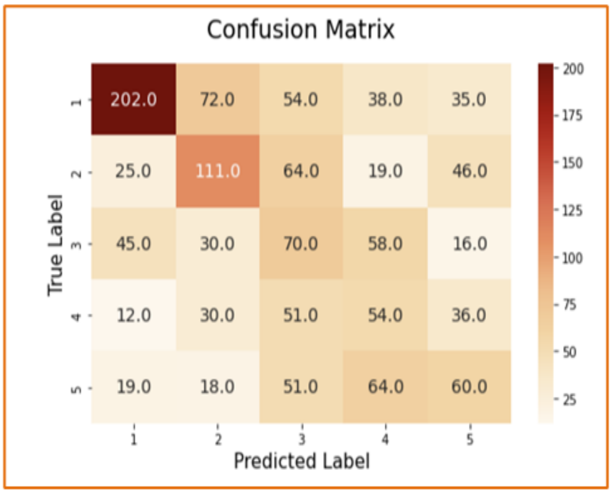
Data set	Tuned Adaboost oversampled	Tuned Gradient oversampled	XG Boost oversampled
Moderately imbalanced	'n_estimators': 800 'max_features': 'auto' 'max_depth': 140	'n_estimators': 1000 'max_features': 'auto' 'max_depth': 120	'n_estimators': 1000 'max_features': 'auto' 'max_depth': 130
Extremely imbalanced	'n_estimators': 600 'max_features': 'auto' 'max_depth': 250	'n_estimators': 600 'max_features': 'auto' 'max_depth': 250	'n_estimators': 800 'max_features': 'auto' 'max_depth': 250

Table 8 Classification models result for each boosting algorithm.

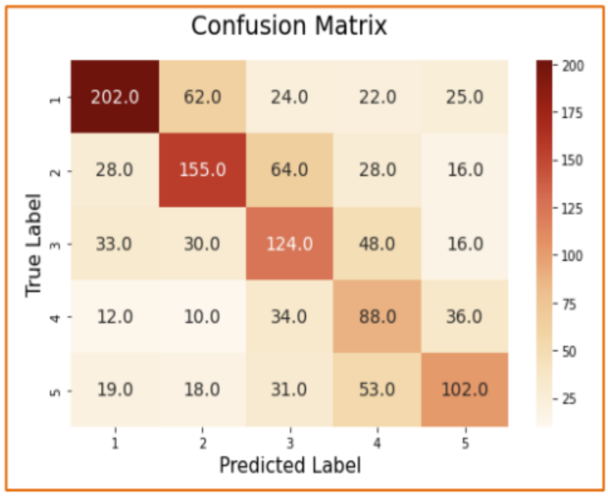
Data set	Performance	AdaBoost		Gradient Boost		XGBoost	
		Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
Default Imbalanced Dataset	Training set	0.675	0.677	0.745	0.747	0.781	0.789
	Test set	0.666	0.659	0.737	0.732	0.765	0.768
Tuned Oversampled Dataset	Training set	0.988	0.986	0.821	0.820	0.791	0.792
	Test set	0.676	0.673	0.765	0.763	0.789	0.788

Fig. 5 shows the confusion matrix of the Gradient Boosting classification models for the test set. The values in confusion matrix showed better prediction than AdaBoost model and resulted in better performance. Gradient boost classifier from scikit-learn package for the default hyperparameter tuning was adopted. Randomized hyperparameter tuning with oversampling technique was used to improve the model performance with accuracy and F1 score 0.765 and 0.763 for test set. This model still showed slightly overfitting when comparing the performance of training set and test set.

Fig. 6 shows the confusion matrix of the XGBoost classification models for the test set. The values in confusion matrix showed better prediction than the previously developed boosting models and resulted in better performance and the model did not seem to be overfitting both in training and test set. Gradient boost classifier from scikit-learn package for the default hyperparameter tuning was adopted to develop the XGBoost model with imbalanced dataset. Randomized hyperparameter tuning with oversampling technique was used to improve the model performance with accuracy and F1 score 0.789 and 0.788 for test set.

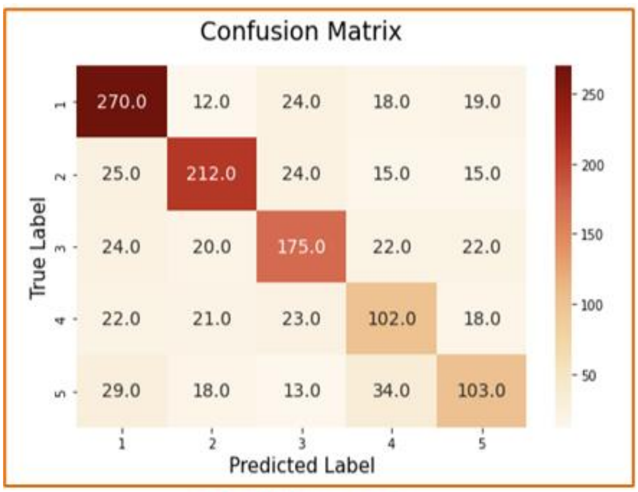


Default Imbalanced Set

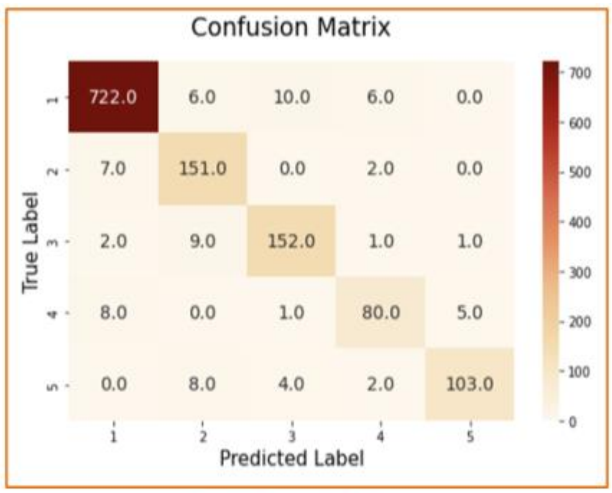


Tuned Oversampled Set

Fig. 4 Confusion matrix for AdaBoost classification models.



Default Imbalanced Set



Tuned Oversampled Set

Fig. 5 Confusion matrix for Gradient Boosting classification models.

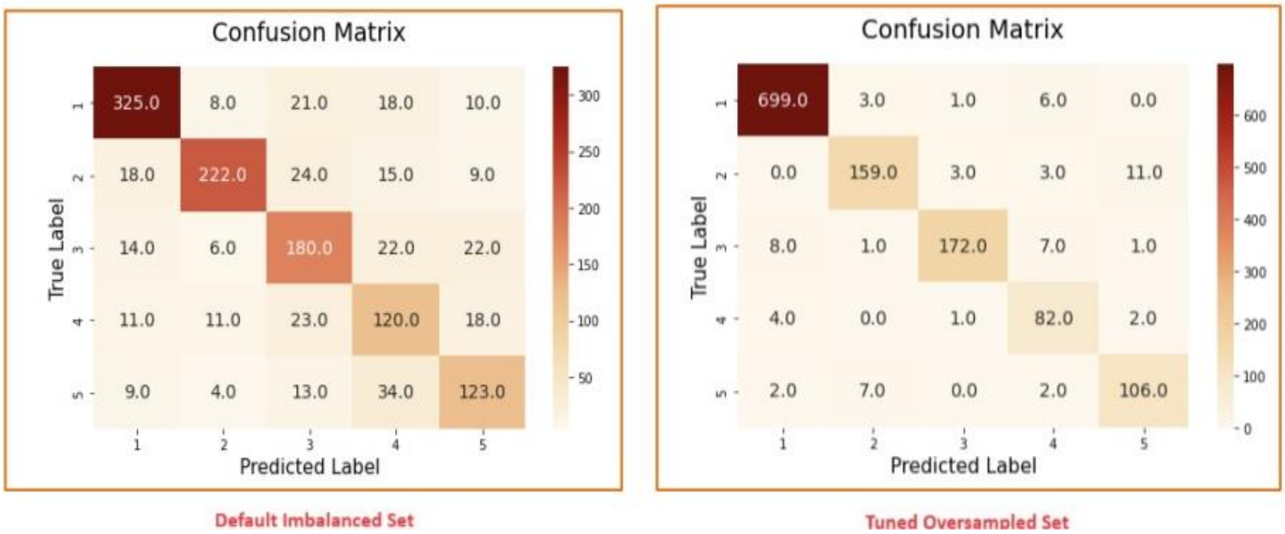


Fig. 6 Confusion matrix for XGBoost classification models.

6. Discussions

In this study, there are a total of 6 tree-based models that were developed to evaluate and compare the performance of each model to choose the best model to identify the critical variables in the data collected. Fig. 7 summarizes the test performance of all the tree-based boosting models.

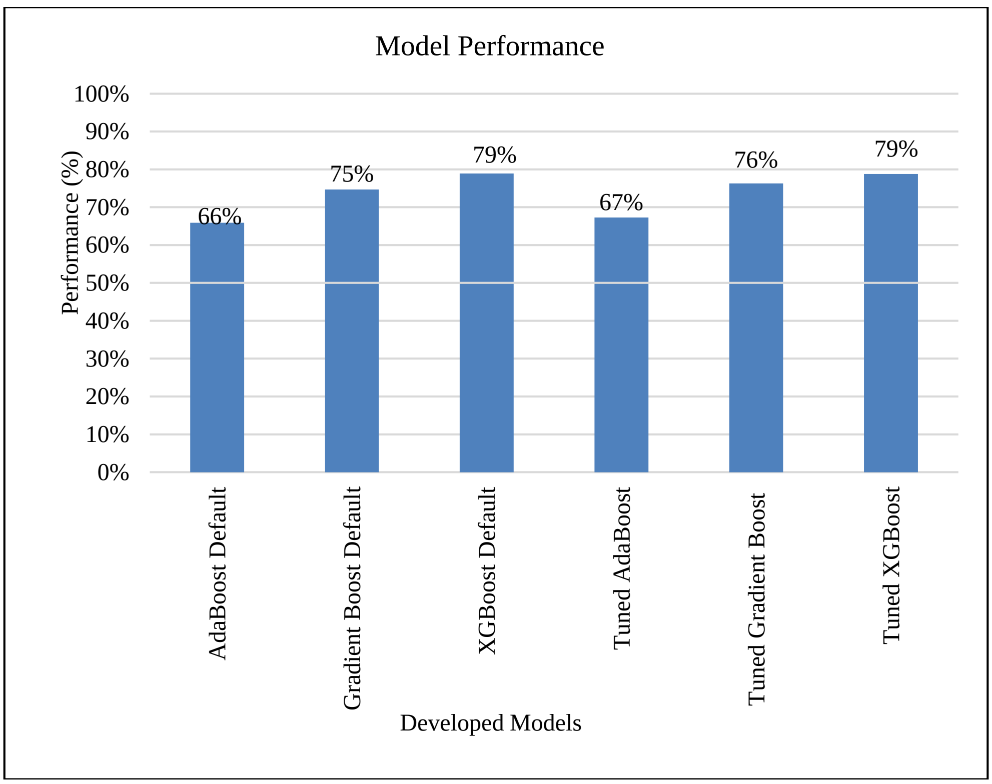


Fig. 7 Model performance.

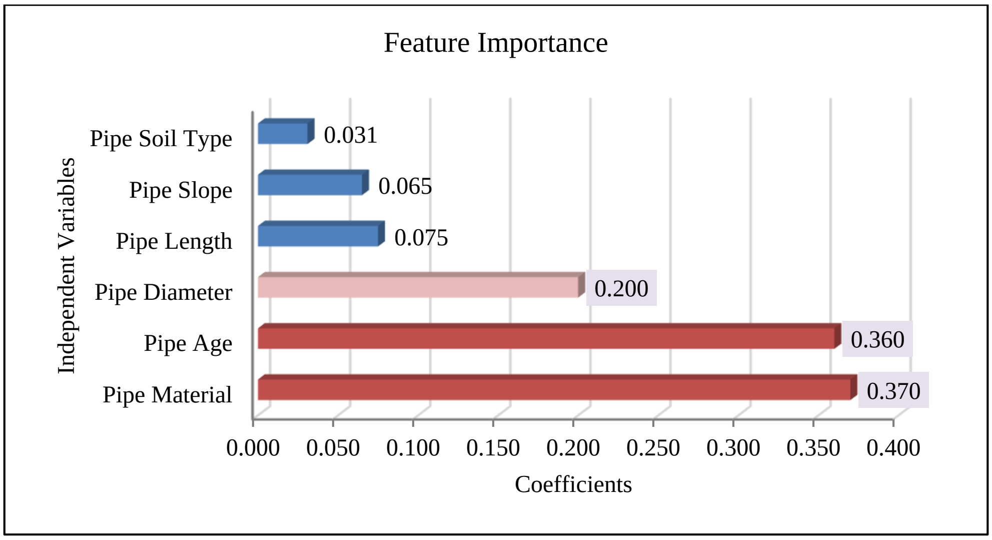


Fig. 8 Feature importance of tuned XGBoost Model.

As the test performance of tuned XGBoost algorithm is better than all the other models, the feature importance of the variables for this model was developed and is presented Fig. 8. It can be seen that pipe material and pipe age exhibit the high impact on the PACP score with the relative importance of 0.36 and 0.37 and pipe diameter is the second most impacting variable with the co-efficient of 0.20. The pipe length and slope have low scores of 0.035 and 0.028. Finally pipe soil type has the lowest score of 0.0015 and it shows no impact on the PACP score. In order to evaluate on this conclusion, XGBoost model was reconstructed to check its performance by including the pipe age, pipe material, pipe diameter and pipe length variables in the dataset and it was found that, the macro F1-score for this model was 0.80 and this model has outperformed than the previous developed model with all the independent variables. Therefore, it can be decided that the critical variables mentioned above in this section can be considered to have more effect on the PACP score by neglecting pipe slope and pipe native soil type.

7. Conclusions

It is well understood that inspecting each separate sanitary sewer pipe segment is not achievable for any municipality because of the enormous inventory of pipes and associated costs. ML boosting algorithms were employed to develop condition prediction models that could point on the sanitary sewer lines which need repair or maintenance. Also, it was evident that the sewer inspection data were imbalanced, and it has to be dealt with random oversampling techniques in order to create a more effective prediction model otherwise would result in the model with questionable results. This study compared the performance of default imbalanced models and tuned oversampled models by developing the ensembled boosting algorithms namely—AdaBoost, Gradient boost and XGBoost methods. Our goal was to demonstrate the reduction of class imbalance in the sewer pipe dataset with varying imbalance ratios. This paper can help future researchers to understand the benefits of resampling methods that address data imbalances. The models developed can help utility owners and municipal assets managers make better decisions about future sewer network inspections.

8. Limitations

This research is undertaken to demonstrate the prediction model developed using tree-based ML algorithms which are expected to produce better model performance. The main limitation of condition prediction models is the availability of appropriate datasets to generate the models. Environmental parameters affecting the condition of sanitary sewer pipes, such as bedding material, overburden pressure, soil water content, traffic flow, and other factors identified in the literature, were omitted due to a lack of a proper dataset. On the other hand, the population of sanitary sewer pipes in condition levels 1, 2, and 3 was found to be more in number in the given dataset for the region with respect to levels 4 and 5. This variable in the dataset caused low performance in developed models like random forest, which has always proved to be one of the efficient approaches to build a prediction model.

References

[1] Najafi, M., and Gokhale, S. B. 2022. *Trenchless Technology: Pipeline and Utility Design, Construction, and Renewal*. New York: McGraw-Hill.

[2] Arjun, M. S., Rahman, S., Nanjundappa, A., and Hammouri, A. P. 2024. "Adapting Oversampling and Hyperparameters Tuning to Improve the Performance of Prediction Model for Sewers." In *Proceedings of the ASCE Pipeline Conference*, July 27-31, 2024, Calgary, Alberta, Canada.

[3] EPA. 2007. "Innovation and Research for Water Infrastructure for the 21st Century." www.Ep.gov/nrmrl.

[4] EPA. 2009. "Condition Assessment of Wastewater Collection Systems." www.Ep.gov/nrmrl.

[5] ASCE. 2021. *Report Card for America's Infrastructure*. New York: American Society of Civil Engineers (ASCE).

[6] American Water Works Association (AWWA). 2012. "Buried No Longer: Confronting America's Water Infrastructure Challenge." https://www.awwa.org/Portals/0/files/legreg/documents/BuriedNoLonger.pdf.

[7] Atambo, D., Rahman, S., Arjun, M., Kaushal, V., Tehrani, A., and Najafi, M. 2023. "Factors Influencing Watermain Break Rates." In *Proceedings of the ASCE Pipeline Conference*, August 12-16, 2023, San Antonio, TX, U.S.

[8] Rahman, S., Arjun, M., Atambo, D., Hassi, S., and Jibreen, A. 2023. "Municipal Pipe Material Joints: Design and Sealing." In *Proceedings of the ASCE Pipeline Conference*, August 12-16, 2023, San Antonio, TX, U.S.

[9] EPA. 2015. "Condition Assessment of Underground Pipes." www.Ep.gov/nrmrl.

[10] Arjun, M., Kaushal, V., Shirkhanloo, S., Rahman, S., and Najafi, M. 2023. "A Review of Rehabilitation Methods for Aging Pipe Culverts by Spray Applied Pipe Lining, Grouting, and Sliplining." In *Proceedings of the ASCE Pipeline Conference*, August 12-16, 2023, San Antonio, TX, U.S.

[11] Geisbush, J., Collins, C., and Arjun, M. P. 2024. "Developing and Executing Large Diameter Water Pipeline Physical Entry Inspection Protocols." In *Proceedings of the ASCE Pipeline Conference*, July 27-31, 2024, Calgary, Alberta, Canada.

[12] Opila, M. C. 2011. "Structural Condition Scoring of Buried Sewer Pipes for Risk-Based Decision Making." Ph.D. thesis, University of Delaware.

[13] National Association of Sewer Service Companies (NASSCO), Pipeline Assessment Certification Manual, Dallas (2015).

[14] Shirkhanloo, S., Arjun, M., Najafi, M., Kaushal, V., Kaur, K., and Jibreen, A. 2023. "Investigation of KNN and Decision Tree Methods' Efficiency in Developing Prediction Models for Sewer Pipes." In *Proceedings of the ASCE Pipeline Conference*, August 12-16, 2023, San Antonio, TX, U.S.

[15] Najafi, M., and Kulandaivel, G. 2005. "Pipeline Condition Prediction Using Neural Network Models." Pipelines 2005.

[16] Chughtai, F. R. M. 2008. "Integrated condition assessment models for sustainable sewer pipelines." thesis

[17] Mashford, J., Marlow, D., Tran, D., and May, R. (2011). "Prediction of Sewer Condition Grade Using Support Vector Machines." Journal of Computing in Civil Engineering, 25(4), 283–290.

[18] Sousa, V., Matos, J. P., and Matias, N. 2014. "Evaluation of artificial intelligence tool performance and uncertainty for predicting sewer structural condition." Automation in Construction, 44, 84–91.

[19] Kabir, G., Balek, N. B. C., and Tesfamariam, S. 2018. "Sewer Structural Condition Prediction Integrating Bayesian Model Averaging with Logistic Regression." Journal of Performance of Constructed Facilities, 32(3), 04018019.

[20] Arjun, M. 2023. "Comparison of Condition Prediction Models to Prioritize Sewer Pipe Inspections." Ph.D. thesis, University of Texas at Arlington.

[21] Ariaratnam, S. T., El-Assaly, A., and Yang, Y. 2001. "Assessment of Infrastructure Inspection Needs Using Logistic Models." Journal of Infrastructure Systems, 7(4), 160–165.

[22] Tran, D., Ng, A., and Perera, B. (2006). "Neural networks deterioration models for serviceability condition of buried stormwater pipes." Engineering Applications of Artificial Intelligence, 20(8), 1144–1151.

[23] Tran, H. D. 2007. "Investigation of Deterioration Models for Stormwater Pipe Systems." Doctoral Dissertation. Victoria University. Melbourne, Australia.

[24] Chughtai, F., and Zayed, T. 2008. "Integrating WRc and CERIU Condition Assessment Models and Classification Protocols for Sewer Pipelines." Journal of Infrastructure

Systems, 17(3), 129–136

[25] Tran, D. H., Perera, B. J. C., and Ng, A. 2009. “Comparison of Structural Deterioration Models for Stormwater Drainage Pipes.” Computer-Aided Civil and Infrastructure Engineering, 24(2), 145–156.

[26] Khan, Z., Zayed, T., and Moselhi, O. 2010. “Structural Condition Assessment of Sewer Pipelines.” Journal of Performance of Constructed Facilities, 24(2), 170–179.

[27] Lubini, A. T., and Fuamba, M. 2011. “Modeling of the deterioration timeline of sewer systems.” NRC Research Press.

[28] Mashford, J., Marlow, D., Tran, D., and May, R. 2011. “Prediction of Sewer Condition Grade Using Support Vector Machines.” Journal of Computing in Civil Engineering, 25(4), 283–290.

[29] Salman, B., and Salem, O. (2012). “Modeling Failure of Wastewater Collection Lines Using Various Section-Level Regression Models.” Journal of Infrastructure Systems, 18(2), 146–154.

[30] Syachrani, S., Jeong, H. S. “D., and Chung, C. S. 2013. “Decision Tree–Based Deterioration Model for Buried Wastewater Pipelines.” Journal of Performance of Constructed Facilities, 27(5), 633–645.

[31] Harvey, R. R., and Mcbean, E. A. (2014). “Predicting the structural condition of individual sanitary sewer pipes with random forests.” Canadian Journal of Civil Engineering, 41(4), 294–303.

[32] Bakry, I., Alzraiee, H., Kaddoura, K., Masry, M. E., and Zayed, T. (2016). “Condition Prediction for Chemical Grouting Rehabilitation of Sewer Networks.” Journal of Performance of Constructed Facilities, 30(6), 04016042.

[33] Gedam, A., Mangulkar, S., and Gandhi, B. 2016. Prediction of sewer pipe main condition using the linear regression approach. Journal of Geoscience and Environment Protection, 100-105.

[34] Hernandez, N., Caradot, N., Sonnenberg, H., Rouault, P., and 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.

[35] Laakso, T., Kokkonen, T., Mellin, I., and Vahala, R. 2018. Sewer Condition Prediction and Analysis of Explanatory Factors. Water, 1239.

[36] Malek Mohammadi, M. 2019. “Development of Condition Prediction Models for Sanitary Sewer Pipes.” Ph.D. thesis, University of Texas at Arlington.

[37] Laakso, T., Kokkonen, T., Mellin, I., and Vahala, R. 2018. Sewer Condition Prediction and Analysis of Explanatory Factors. Water, 1239.

[38] Loganathan, K. 2021. Development of a Model to Prioritize Inspection and Condition Assessment of Gravity Sanitary Sewer Systems (Doctoral dissertation, The University of Texas at Arlington).

[39] Shirkhanloo. S. 2022. Evaluation of Decision-Making Prediction Models for Sewer Pipes Asset Management (Doctoral dissertation, The University of Texas at Arlington).

[40] Müller, A. C., and Guido, S. 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. Boston: O’Reilly.

[41] Salman, B. 2010. “Infrastructure Management and Deterioration Risk Assessment of Wastewater Collection Systems.” Ph.D. thesis, University of Cincinnati.

[42] Teh, K., Armitage, P., Tesfaye, S., Selvarajah, D., and Wilkinson, I. D. 2020. “Imbalanced Learning: Improving Classification of Diabetic Neuropathy from Magnetic Resonance Imaging.” *PLoS One* 15 (12): e0243907.

[43] Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. 2013. *Applied Logistic Regression*. Hoboken, NJ: Wiley.

[44] Hossin, M., and Sulaiman, M. 2015. “A Review on Evaluation Metrics for Data Classification Evaluations.” *International Journal of Data Mining and Knowledge Management Process* 5: 1-11.