International Journal of Advance Research in Science and Engineering
Volume No. 13, Issue No. 09, September 2024
www.ijarse.com
IJARSE
ISSN 2319 - 8354

Risk-Based Decision-Making for Pipe Leakage Detection

Sai Nethra Betgeri[1] Naga Parameshwari Chekuri[2]

[1]Ph.d, Department of Computer Science, University of Louisville, Duthie Center of Engineering, 208, Louisville, KY, 40245.

[1]sainethra.betgeri@louisville.edu, [1]sainethra.betgeri@gmail.com

[2]Ph.d Department of Management Studies, Dr. B. V. Raju Institute of Technology, Narsapur, Telangana, 502313

[2]parameshwari.chekuri@bvrit.ac.in

ABSTRACT

Risk-based assessments of pipe conditions focus on prioritizing critical assets by evaluating the risk of pipe failure. The aging wastewater infrastructure is a growing concern for utilities across the country. The US water sector received a concerning C- grade (Report, 2021), an improvement from a previous D, while the wastewater sector earned a troubling D+ in the latest Infrastructure Report Card. Over the next 25 years, $271 billion will be required to maintain and operate these networks effectively. Furthermore, the demand for wastewater collection and treatment is projected to increase by 23% by 2032. However, leaks in wastewater pipelines are a major source of loss for operators, potentially causing severe ecological disasters, human casualties, and financial loss. Traditional manual methods for assessing structural leakage in sewage pipes are time-consuming. This study introduces an automated method using K-Nearest Neighbors (K-NN) to effectively identify pipe leaks using repair data. This classification process helps to quickly identify wastewater pipes needing immediate replacement. The proposed model is tested on a Phase-3 US wastewater collection system in Shreveport, Louisiana.

Keywords: Risk, *K*-Nearest Neighbors (*K*-NN), Automation.

1. INTRODUCTION

The aging wastewater infrastructure is an increasing concern for utilities across the country. In the 2021 Infrastructure Report Card, the US water sector received a concerning C- grade, a slight improvement from its previous D rating (EPA, 2004), while the wastewater sector earned a D+. Over the next 25 years, an estimated $271 billion will be required to maintain and manage these systems at an adequate operational level. Additionally, demand for wastewater collection and treatment is projected to grow by 23% by 2032 (Report, 2021). Risk-based asset management focuses on identifying the most critical assets to determine the most effective strategies for detecting pipe leaks, rehabilitating, and replacing pipe infrastructure. The Pipeline Assessment and Certification Program (PACP), established by the National Association of Sewer Service Companies, is the industry-standard protocol for assessing and managing the condition of sewer pipes in the United States.

The Pipeline Assessment and Certification Program (PACP), established by the National Association of Sewer Service Companies, is the industry-accepted and used protocol for rehabilitation and replacement of the condition of sewer pipes in the United States (Angkasuwansiri & Sinha, 2015; Aprajita, 2018; Betgeri, 2022; Betgeri,

Matthews, et al., 2023; Betgeri et al., 2024; Betgeri, Vadyala, et al., 2023; DeBoda & Bayer, 2015). Since the initial development of the method, several updated versions exist, the most current one is PACP version 7.0.4, released on October 1, 2020. The PACP method relies entirely on visual inspections using closed-circuit television (CCTV). Trained operators assess structural and operation and maintenance (O&M) issues. A CCTV camera is mounted on an IBAK crawler with a 1000-foot cable, transmitting high-resolution images to a computer and display above ground. As the crawler moves through the pipe, continuous video is recorded. The crawler can be paused at any point, allowing the CCTV camera to rotate and zoom in on areas of interest for more detailed inspection. The inner surface of the pipe is recorded in real-time during the inspection, and contractors analyze the footage immediately. Based on the CCTV inspections, contractors generate pipe assessment reports, and inspectors classify pipe failures according to the industry-standard PACP protocol for all reports. Based on the pipe leak failures classification maintenance is scheduled. The overall leak detection protocol is shown in Figure 1. Figure 2 shows the pipe leakage in a wastewater pipe.

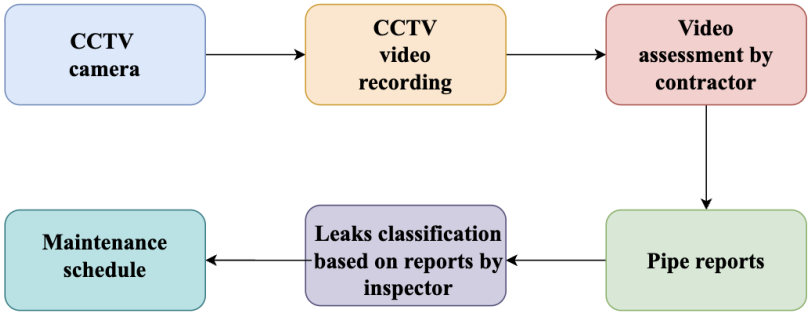

Figure 1: Overall leak detection protocol

Figure 2: Pipe leakage in a wastewater pipe

2. OBJECTIVE

The main objective of this paper is to automate the leak classification based on the reports by inspectors to classify pipe failures and schedule maintenance faster.

3. METHODOLOGY

3.1 DATASET

A total of 3100 pipe data totaling approximately 285 km (935,703 ft) is given. For this study, a total length of roughly 47 km (154,060 ft) of 200 mm (8 in.) diameter vitrified clay (VC) pipe, totaling 3100 pipe segments, was selected. Information such as Pipe ID, Leaks observed or not is mentioned in the pdf format. The pdf data is given by the Dept. of Engineering & Environmental Services, Shreveport, Louisiana Phase 3. We used Python programming to process the records of all the PDF documents into a CSV file. Figure 3 shows the data extraction process.

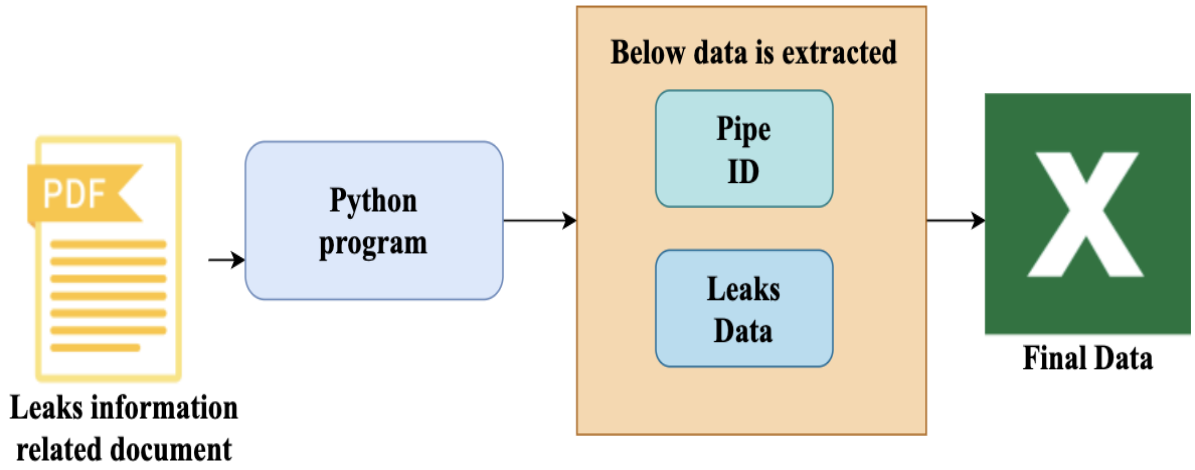

Figure 3: Data extraction

3.2 DATA PREPROCESSING

Data Preprocessing is when the data gets transformed, or encoded, such that the machine can quickly parse it. In this study, we included records with relevant data by removing inconsistent data, and missing information info per pipe for further analysis. This step makes the training dataset cleaner and error-free, which helps in improving the accuracy of the model. After all these analyses and verification of data, the final data collection included 2970 pdf reports for our analysis as shown in Fig. 4.

3.2.1 MISSING VALUES

It is very usual to have missing values in our dataset. It may have happened during data collection by the CCTV inspector. We eliminated 60 reports related to the few missing information related to leaks observed.

3.2.2 INCONSISTENT VALUES

We know that data can contain inconsistent values. Due to human error, or maybe the information was entered as not sure about leaks observed. We have eliminated 70 reports related to the inconsistent values.

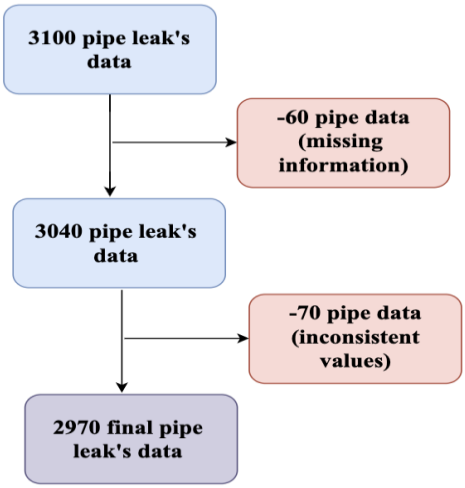

Figure 4: Process of leaks final data

3.3 LEAK DETECTION FRAMEWORK

The leak detection framework model incorporates the well-established industry-standard condition rating method, the PACP, developed by NASSCO in 2014. The *K*-Nearest Neighbor model is used for this purpose. *K*-NN is a non-parametric method used for classification. The basic logic behind *K*-NN is to explore your neighborhood, assume the test data point to be like them, and derive the output. Compared to other classifier algorithms, it is very easy to implement. If training data is much larger than several features (m≫n), *K*-NN is better than SVM. Compared to Neural networks, it requires less training data to achieve the same accuracy.

We didn't consider the geographical location of the pipe for our model implementation. A leak detection framework is shown in Fig. 5.

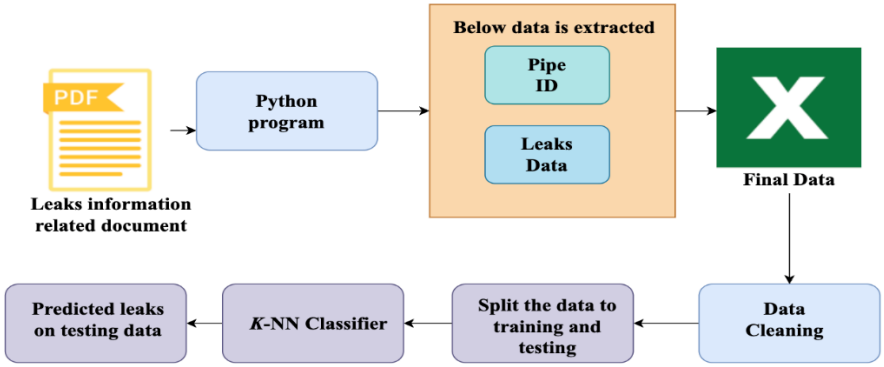

Figure 5: Leak detection framework

3.3.1 *K* −nearest Neighbor (*K*-NN)

The *K*-nearest neighbor's algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. (Peterson, 2009). *K*-NN classifies the new data points based on the similarity measure of the earlier stored data points.

Compared to other algorithms *K*-NN is called Lazy Learner (Instance-based learning). It does not learn anything in the training period. It does not derive any discriminative function from the training data. It stores the training dataset and learns from it only at the time of making real-time predictions. This makes the *K*-NN algorithm much faster than other algorithms that require training e.g., SVM, Linear Regression, etc. New data can be added seamlessly at any point in time which will not impact the accuracy of the algorithm. Finally, it is very easy to implement because it only requires two parameters *K* and the Euclidean distance function.

Algorithm:

Input: E: All factors, K: Chosen Number of Neighbors

Output: C: Mode of K labels

Begin:

- Load the data.
- Initialize *K* to your chosen number of neighbors.
- For each testing data:
  - Calculate the distance between 25% of testing data (x, y) with all 75% of the training data. (a, b) using Euclidean distance (ED) as shown in Equation 2.

$$ED = \sqrt{(x-a)^2 + (y-b)^2} \qquad (Eq.1)$$

  - Add the distance and the index of testing data to the ordered collection.
- Sort the ordered collection of distances and indices in ascending order by distances.
- Pick the first *K* entries from the sorted collection.
- Get the labels of selected entries.
- Return the mode of *K* labels.

End

4. RESULTS AND ANALYSIS

We have divided the data into 75% training and 25% validation data, and the process is repeated several times with different values of K to reduce the errors and to make accurate predictions. We have finally chosen the value as $K = 9$. As the value of K is increased, our predictions become more stable and will have more accurate predictions up to a certain point. Figure 7 shows the graph of the misclassification rate as a function of *K* for 20 and 25, and from both graphs, we see the lowest error is found at $K = 9$ with a value of 0.012. We also checked for different values of *K*, and we found the lowest value of the misclassification rate at 9. So, we have used the value as $K = 9$ for better accuracy. Table 1 shows the count and misclassification rate for training data and testing data for *K=20* and Table 2 shows the count and misclassification rate for training and testing data for *K=25*. Figure 6 shows the plot of the misclassification rate for $K = 20$. Figure 7 shows the plot of the misclassification rate for $K = 20$. Misclassification can be reduced when the model is trained with a wider variety of data.

Table 1: Misclassification rate for *K=20*

	Training		Testing	
K	Count	Misclassification Rate	Count	Misclassification Rate
1	2227	0.194	743	0.184

2	2227	0.187	743	0.177
3	2227	0.165	743	0.165
4	2227	0.135	743	0.145
5	2227	0.102	743	0.112
6	2227	0.082	743	0.092
7	2227	0.065	743	0.057
8	2227	0.034	743	0.032
9	2227	0.012	743	0.011
10	2227	0.015	743	0.018
11	2227	0.023	743	0.025
12	2227	0.030	743	0.032
13	2227	0.060	743	0.059
14	2227	0.045	743	0.047
15	2227	0.035	743	0.039
16	2227	0.055	743	0.064
17	2227	0.084	743	0.075
18	2227	0.123	743	0.114
19	2227	0.186	743	0.192
20	2227	0.176	743	0.183

Table 2: Misclassification rate for *K=25*

	Training		Testing	
K	Count	Misclassification Rate	Count	Misclassification Rate
1	2227	0.185	743	0.173
2	2227	0.177	743	0.167
3	2227	0.155	743	0.162
4	2227	0.145	743	0.153
5	2227	0.112	743	0.132
6	2227	0.092	743	0.083
7	2227	0.055	743	0.049
8	2227	0.044	743	0.021
9	2227	0.013	743	0.011
10	2227	0.016	743	0.025
11	2227	0.019	743	0.029
12	2227	0.025	743	0.032
13	2227	0.055	743	0.063
14	2227	0.040	743	0.045
15	2227	0.032	743	0.038
16	2227	0.060	743	0.057
17	2227	0.074	743	0.077
18	2227	0.090	743	0.092
19	2227	0.102	743	0.112
20	2227	0.135	743	0.124
21	2227	0.142	743	0.134
22	2227	0.153	743	0.159
23	2227	0.167	743	0.170
24	2227	0.177	743	0.175
25	2227	0.165	743	0.163

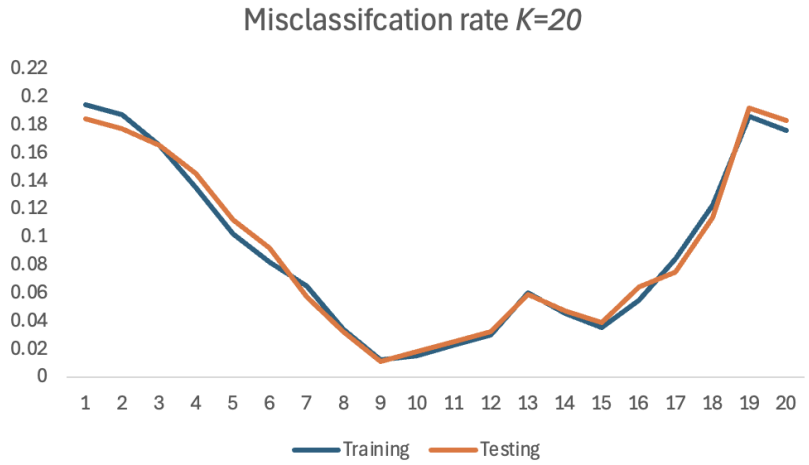

Figure 6: Misclassification rate for *K*=20

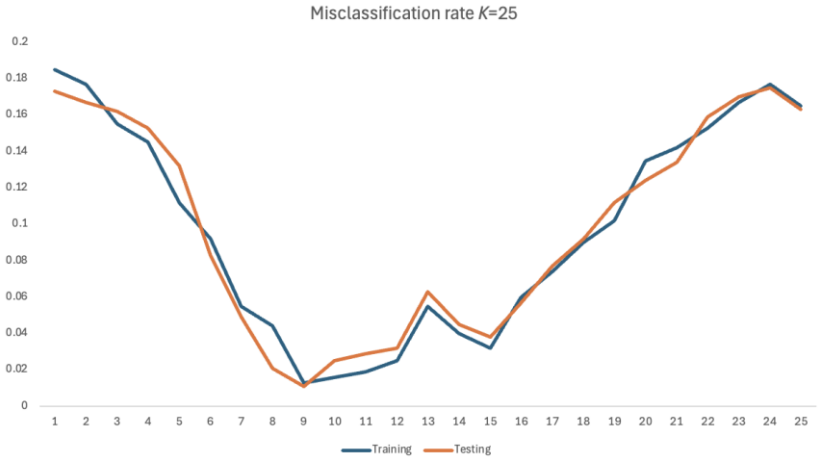

Figure 7: Misclassification rate for *K*=25

To proceed with the *K*-NN calculation process, Euclidian distance is used to find the distance between each testing data to training data as shown in Equation 2. Table 3 shows the confusion matrix of validation data compared with the original result given by the inspector. Table 4 shows the accuracy, precision, recall, and F1 score for leaked data, and Equations 2 through 5 present the accuracy, precision, recall, and F1 score, respectively.

$$Accuracy = \left(\frac{TP+TN}{TP+TN+FP+FN}\right)*100\% \quad (Eq.2)$$

$$Precision = \frac{TP}{TP+FP} \quad (Eq.3)$$

$$Recall = \frac{TP}{TP+FN} \quad (Eq.4)$$

$$F1\ Score = \frac{2TP}{2TP+FP+FN} \quad (Eq.5)$$

where TP, FN, FP, and TN represent the number of true positives, false negatives, false positives, and true negatives, respectively. In summary, the *K*-NN classifiers are based on leaks data will reduce the manual efforts of the inspector.

Table 3: Confusion matrix

	Actual Leaks	Actual No Leaks
Predicted Leaks	385	6
Predicted No Leaks	8	344

Table 4: Accuracy, precision, recall, and F1 score

Accuracy	0.9812
Precision	0.9847
Recall	0.9796
F1 Score	0.9821

5. CONCLUSION

The proposed leak detection framework assesses the leaks in the wastewater pipe in must faster way by saving the pipe from more deterioration. A *K*-Nearest Neighbor (*K*-NN) model was used to automate the pipe leaks reduce the efforts of the inspector and speed up the process. To validate the model, the predicted leak detection of our model was compared with the actual leak classification given by the inspector, and our accuracy was 98.12% which is satisfactory.

One of the main limitations of the study was the data. Therefore, more pipe from different geographic locations is needed to improve and convey more robustness to the obtained results. The other limitation was the execution time because *K*-NN Classifiers are real-time execution, so their execution is slow compared to other classifier algorithms.

REFERENCES

1. Angkasuwansiri, T., & Sinha, S. (2015). Development of a robust wastewater pipe performance index. *Journal of Performance of Constructed Facilities*, *29*(1), 04014042.
2. Aprajita, F. (2018). *Guidelines for Implementing Risk-Based Asset Management Program to Effectively Manage Deterioration of Aging Drinking Water Pipelines, Valves and Hydrants* Virginia Tech].
3. Betgeri, S. N. (2022). Analytic Hierarchy Process is not a Suitable method for the Comprehensive Rating.
4. Betgeri, S. N., Matthews, J. C., & Vladeanu, G. (2023). Development of Comprehensive Rating for the Evaluation of Sewer Pipelines. *Journal of Pipeline Systems Engineering and Practice*, *14*(2), 04023001.
5. Betgeri, S. N., Vadyala, S. R., & Matthews, J. C. (2024). Probability and Consequence of Failure for Risk-Based Asset Management of Wastewater Pipes for Decision Making. *Civil Engineering Research Journal*, *14*(4).
6. Betgeri, S. N., Vadyala, S. R., Matthews, J. C., & Lu, H. (2023). Wastewater pipe defect rating model for pipe maintenance using natural language processing. *Frontiers in Water*, *5*, 1123313.
7. DeBoda, T., & Bayer, J. (2015). Benefits of PACP® version 7.0 update NASSCO. In *Pipelines 2015* (pp. 878-886).
8. Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883.