Advanced-Data Analytics for Water Loss Management and Leakage Detection using Machine Learning Models: A Case Study of Al Seeb Area, Khoudh Six

Ibrahim Al Balushi, Muzammil Hussain*

***Corresponding Author**

Global College of Engineering and Technology, Sultanate of Oman

DOI : https://doi.org/10.51244/IJRSI.2024.1114SDAICO6

Received: 04 July 2024; Accepted: 19 July 2024; Published: 02 November 2024

ABSTRACT

Using a novel machine learning framework, this study addresses water losses in urban water distribution networks. The focus is on the CatBoost classifier, which predicts leaks with 99.6% accuracy. Using a curated dataset and multiple ML algorithms, the study discovers important relationships that influence leak detection and increase operational efficiency. Despite challenges such as overfitting and limited data hindering broader applicability, the results highlight the need for diverse data and hybrid models. Future work aims to refine these models through real-world testing and assess their environmental and economic impacts.

Keywords: component, CatBoost; Decision Tree, K- Nearest Neighbour, Naïve Bayes, Random Forest, Support Vector Machines.

INTRODUCTION

The growing human population, particularly in urban areas, poses increasing challenges in managing water distribution systems (WDNs) due to increased demand, necessitating more efficient water loss control. The main causes of water loss in these systems are leaks and burst pipes that occur between water treatment plants and distribution to consumers [1]. Water utilities responsible for maintaining WDNs face growing financial burdens due to the need to install new pipelines and maintain aging infrastructure [2]. These challenges are compounded by the high operational costs associated with pumping, treatment, and distribution, leading utilities to develop techniques to detect, locate, and repair leaks [3] more effectively. In fact, about a third of water utilities worldwide report that up to 40% of treated water is lost due to leaks [4]. WDNs consist of complex networks of often inaccessible underground pipes, making leak detection difficult. Traditional leak detection methods require frequent inspections by maintenance teams that do not allow for real- time monitoring, potentially resulting in significant economic losses and environmental damage if leaks go undetected [5].

This research has advanced efforts to improve leak detection in WDNs by improving the accuracy of detection through analysis of leak histories and hydraulic modeling parameters. This study solution includes the development and implementation of machine learning (ML). framework designed to improve the accuracy and efficiency of leak detection in WDNs. This ML approach uses a set of sophisticated algorithms to analyze data from leakages data distributed throughout the water network. This makes it possible to predict the level of each pipe that needs to be replaced as part of the rehabilitation plans to reduce NRW and maintain healthy WDNs. The paper contains the following sections: Section II provides a summary of the research conducted on this topic. Section III explains the data set, Section IV explains the methodology of the work, Section V examines the different algorithms, Section VI compares the results and results, and Section VII concludes and addresses future developments.

RELATED WORKS

Numerous scientists have written a wealth of articles, studies, and research papers on the topic of life expectancy.

One such study conducted by Liu et al. [6] presented a novel water pipe leak detection system that leverages both machine learning and wireless sensor networks to effectively detect leaks in water pipes. Your system integrates ZigBee and 4G technologies for robust signal capture and transmission. To improve the efficiency and lifespan of the system while saving energy, they implemented a leak-triggered network method. The core of their detection capability lies in the use of Empirical Mode Decomposition (EMD), Approximate Entropy (ApEn), Principal Component Analysis (PCA), and Support Vector Machines (SVM) to intelligently identify leak signatures. This approach demonstrated high effectiveness in simulations and field tests, achieving an impressive 98% accuracy in leak classification.

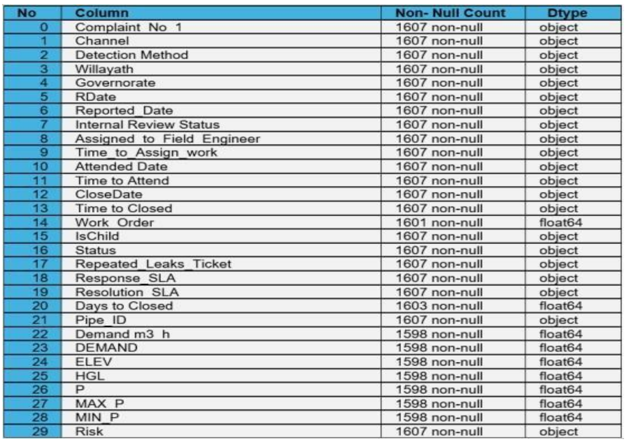
Similarly, Shravani et al. [7] developed an intelligent water management system that applies machine learning techniques to detect and predict leaks in piping systems. Their research highlighted the effectiveness of the Multi-Layer Perceptron (MLP) model, which notably performed with an accuracy of 94.47% and an F1 score of 0.95, demonstrating its reliability in real-world applications. Further contributions to this field came from Alves Coelho, Glória, and Sebastião [8], who examined the use of various machine learning algorithms, including random forest, decision trees, neural networks, support vector machines (SVM), and XGBoost. Their results showed that Random Forest consistently outperformed other models in various scenarios, achieving almost 75% accuracy. This study highlights the potential of machine learning to improve the predictive capabilities and operational efficiency of leak detection systems. Furthermore, Ebisi et al. [9] addressed the critical issue of minimizing non-revenue water (NRW) losses through the use of Internet of Things (IoT) monitoring devices and artificial intelligence (AI). Her research focused on a realistically sized IoT-enabled water transmission system to perform simulated leak tests in complex, underground pipelines typical of industrial infrastructure and smart cities. The study tested three anomaly detection methods: Isolation Forest (iForest), Support Vector Classification (SVC), and a deep learning model using Recurrent Neural Networks-Long Short-Term Memory (RNN- LSTM). These models achieved accuracies of 83.3%, 94.4% and 83.3%, respectively, demonstrating the potential of these technologies to significantly improve the detection and management of water leaks.

The previous analysis shows that a large number of scientists have already conducted research on this topic. A maximum precision of 98% was achieved in their respective works. Predicting leaks more accurately allows us to anticipate the problem more effectively than if the prediction were less accurate. It is plausible that the previous researchers encountered null values or erroneous data, which may have affected their ability to produce a better result. In contrast, we carefully examine and process our dataset before implementing ML techniques, which is responsible for our superior results.

DATASET

A comprehensive dataset on leaks in the Al Khoudh Six Water Network was collected by NWS in 2022 and 2023. The dataset includes 29 common features and a total of 1607 observations, it consists of detailed records of customer and internal leak reports and includes data points such as complaint ID, reporting channel (e.g. mobile app, call center), detection method, and a range hydraulic parameter as listed in Table 1:

Table1: Leakages Dataset



No	Column	Non- Null Count	Dtype
0	Complaint No 1	1607 non-null	object
1	Channel	1607 non-null	object
2	Detection Method	1607 non-null	object
3	Willayath	1607 non-null	object
4	Governorate	1607 non-null	object
5	RDate	1607 non-null	object
6	Reported_Date	1607 non-null	object
7	Internal Review Status	1607 non-null	object
8	Assigned to Field Engineer	1607 non-null	object
9	Time_to_Assign_work	1607 non-null	object
10	Attended Date	1607 non-null	object
11	Time to Attend	1607 non-null	object
12	CloseDate	1607 non-null	object
13	Time to Closed	1607 non-null	object
14	Work Order	1601 non-null	float64
15	IsChild	1607 non-null	object
16	Status	1607 non-null	object
17	Repeated_Leaks_Ticket	1607 non-null	object
18	Response_SLA	1607 non-null	object
19	Resolution SLA	1607 non-null	object
20	Days to Closed	1603 non-null	float64
21	Pipe_ID	1607 non-null	object
22	Demand m3 h	1598 non-null	float64
23	DEMAND	1598 non-null	float64
24	ELEV	1598 non-null	float64
25	HGL	1598 non-null	float64
26	P	1598 non-null	float64
27	MAX P	1598 non-null	float64
28	MIN_P	1598 non-null	float64
29	Risk	1607 non-null	object

METHODOLOGY

After examining the dataset, all attributes were selected based on their completeness and absence of null values. Correlation matrices were used to improve the model by identifying and removing correlated variables. The dataset divided to the size of 70% for train set and 30% for test set with random state to avoid shuffling of datasets. The data was then graphically visualized and various machine learning algorithms were applied, including Naive Bayes (NB), Artificial Neural Networks (ANN), Decision Trees (DT), Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Random Forest (RF) and CatBoost. This approach helped determine the most effective water leak detection algorithm. Finally, compression was performed between the classifiers to determine the most efficient machine learning techniques for our application.

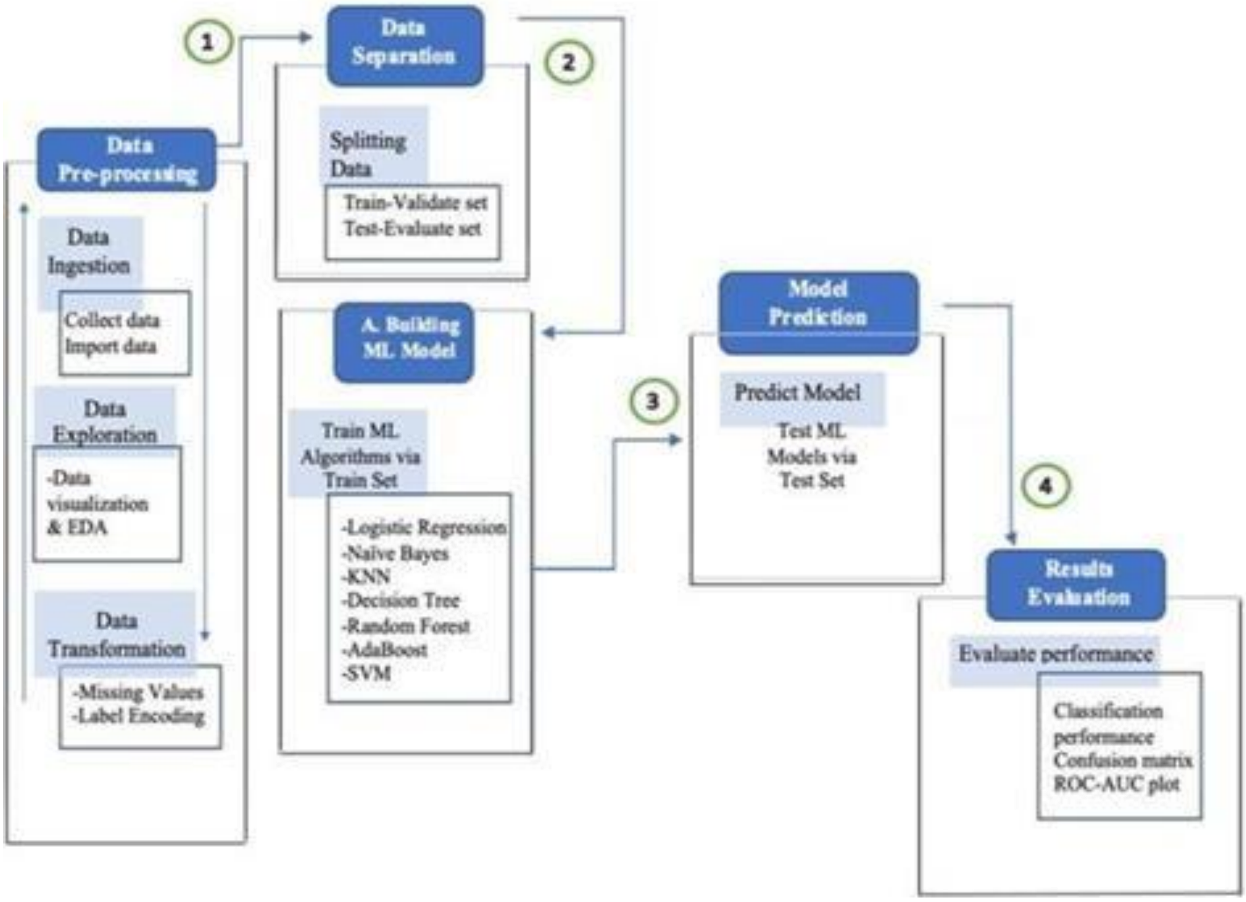


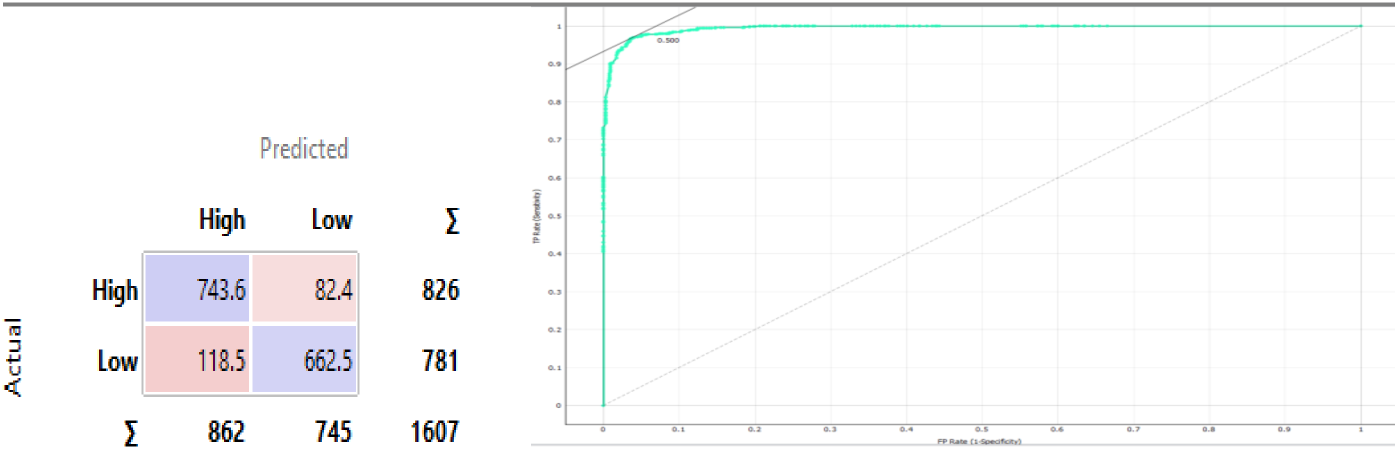
Figure.1. The Flow chart of the methodology used in this research.

MACHINE LEARNING TECHNIQUES

To build the leakage prediction model, the common machine-learning techniques were used as follows:

Random Forset (RF)

The RF technique, a form of supervised ML, is widely used for classification and regression tasks. This approach involves using multiple decision trees to generate an output, hence the name “Random Forest” [10]. However, it has certain limitations when applied to large data sets as it can be computationally intensive and requires more resources than other algorithms. Additionally, overfitting can occur when there is noisy data or datasets with numerous irrelevant features. Another challenge is the difficulty of understanding the results. In Figureure (2) The predicted outcomes are indicated in the confusion matrix that the total numbers of true forecasts are 1406, while the number of incorrect forecasts is 201. It has a training accuracy of 96.5% and testing accuracy of 88.6%. In addition, it has an AUC score of 99.5% and precision of 96.5%. recall of 96.5%, and F1-score of 96.5%.

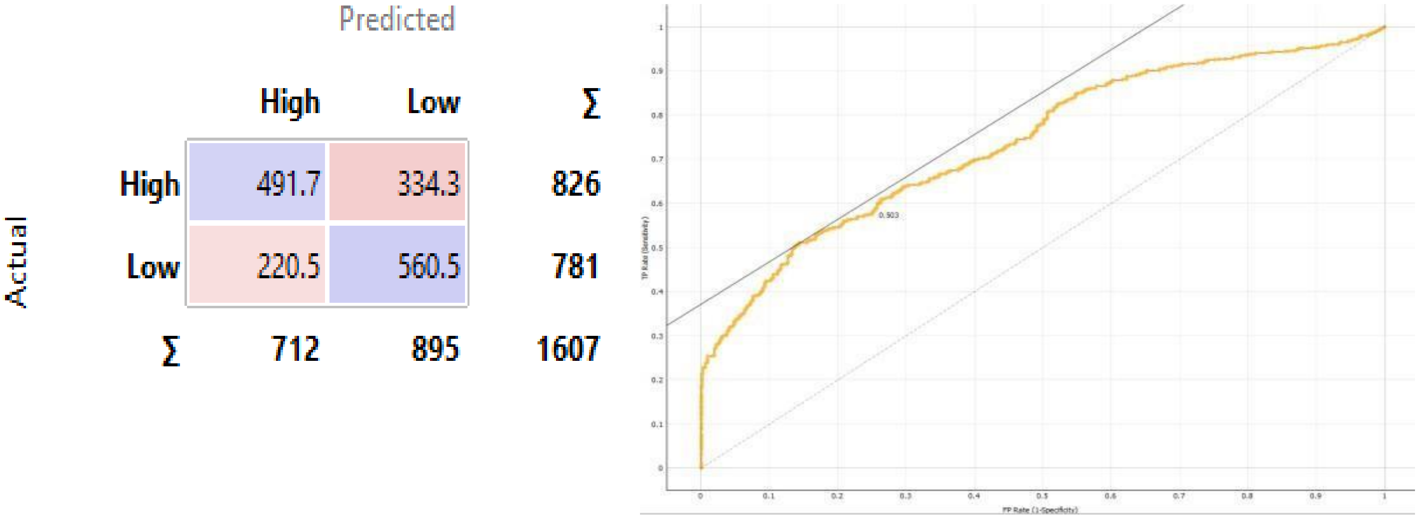


		Predicted		
		High	Low	∑
Actual	High	743.6	82.4	826
	Low	118.5	662.5	781
	∑	862	745	1607

Figure.2. Confusion matrix and ROC plot of RF model.

B.Naive Bayes (NB)

The NB classifier is a type of probabilistic classifier based on Naive Bayes theorem. Its predictive ability depends heavily on the assumption of strong independence between features. This property has made it a popular choice in the field of medical science and disease diagnosis, where it has performed well [11]. The predicted results of the NB model are shown in Figureure (3). The model produces 1052 correct predictions and 555 incorrect predictions, with LR models performing worse than this model in comparison. NB has a training accuracy of 66.5% and a testing accuracy of 66.2%. It also provides a test AUC score, precision, recall and F1 score of 73.9%, 67.1%, 66.5% and 66.3%, respectively.

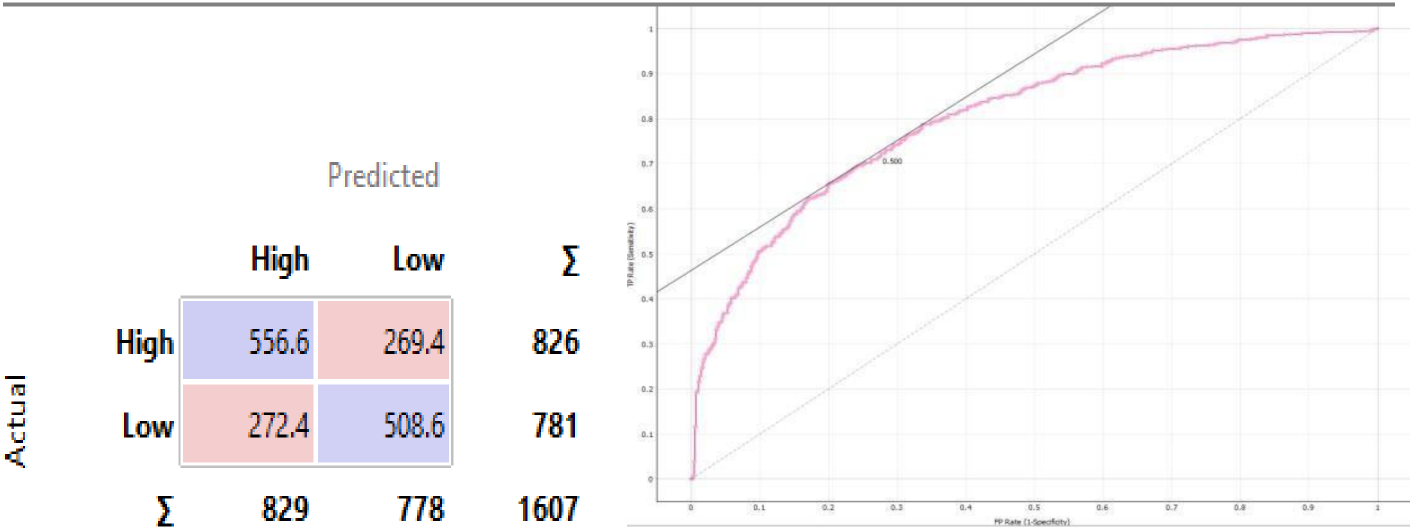


		Predicted		
		High	Low	∑
Actual	High	491.7	334.3	826
	Low	220.5	560.5	781
	∑	712	895	1607

Figure.3. Confusion matrix and ROC plot of NB model.

Artificial Neural Network (ANN)

The algorithm consists of interconnected nodes or neurons, with the output of one node used as input to another node. Each node receives multiple inputs and generates a single output. The multilayer perceptron is a commonly used form of ANN and consists of node layers with an input layer and one or more hidden layers and finally an output layer. [12] The number of neurons present in each layer can vary depending on the circumstances. Figureure (4) shows the prediction of ANN. The predicted results are given in the confusion matrix, so the total number of true predictions is 1065 while the number of false predictions is 542. The training accuracy is 72.3% and the testing accuracy is 66.8%. Furthermore, it has an AUC value of 99.5% and the rest of the performance results are consistent with the accuracy value.

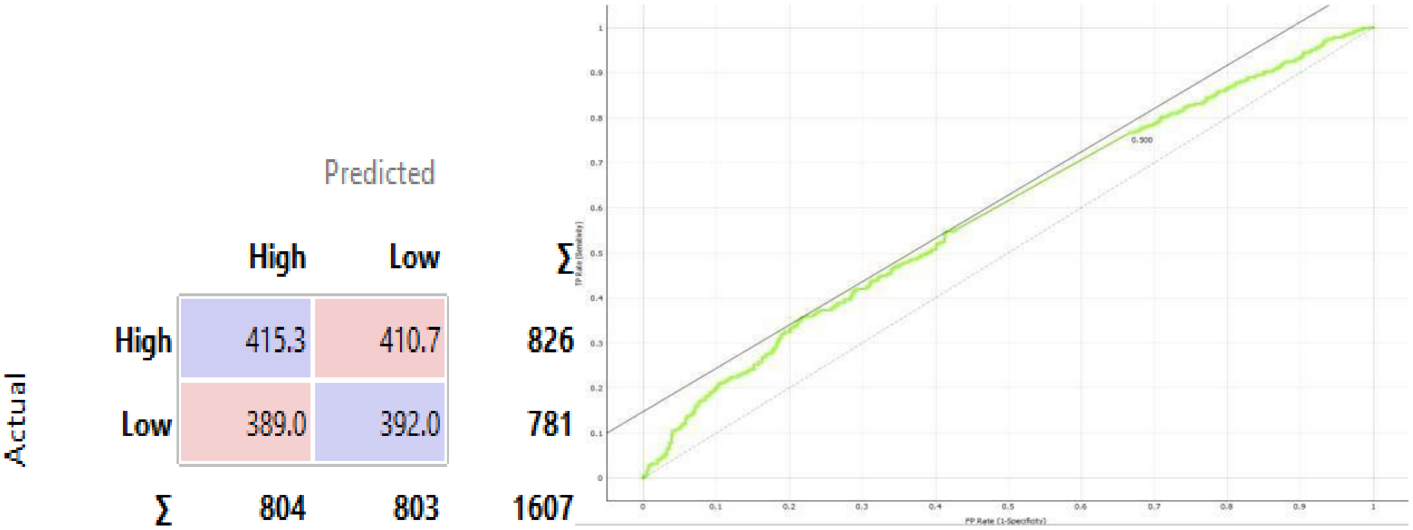


Actual		Predicted High	Predicted Low	∑
	High	556.6	269.4	826
	Low	272.4	508.6	781
	∑	829	778	1607

Figure. 4. Confusion matrix and ROC plot of NN model.

Support Vector Machine (SVM)

SVM is a highly effective discriminative classifier that uses a decision boundary or hyperplane to distinguish between different class labels. The SVM approach identifies the optimal hyperplane from a set of possible options. The ideal hyperplane is determined by maximizing the margin on both sides of the decision boundary. The edge refers to the absence of any data points between the hyperplane and the edge [13]. Figureure (5) shows the result of the final model SVM. The training accuracy is 58.4% and the testing accuracy is 58.2%, which is close to the accuracy of ANN. The AUC score, precision and F1 score are 58.1%, 60.2% and 57.2%, respectively.

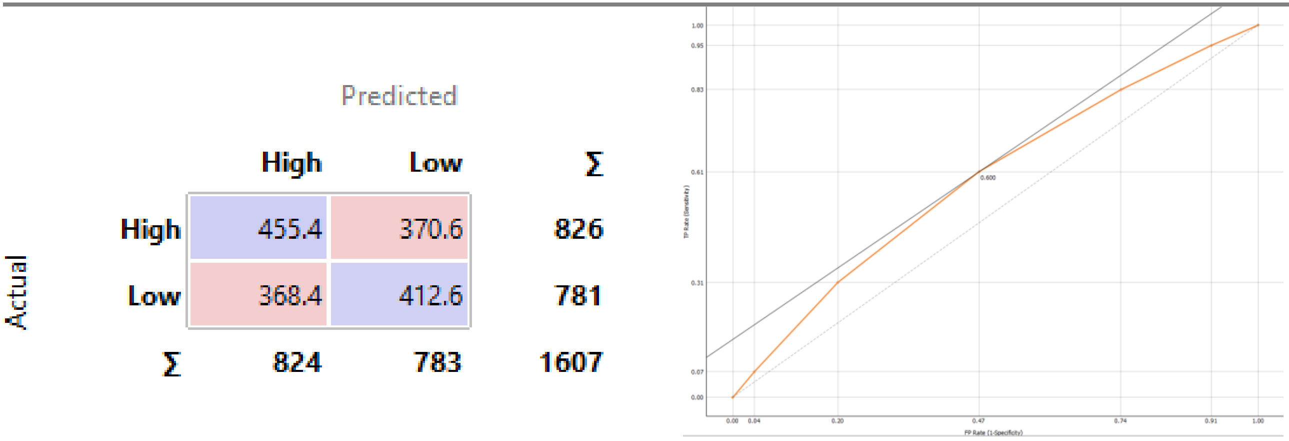


Actual		Predicted High	Predicted Low	∑
	High	415.3	410.7	826
	Low	389.0	392.0	781
	∑	804	803	1607

Figure. 5. Confusion matrix and ROC plot of SVM model.

K-Nearest Neighbor (KNN)

The KNN algorithm is used to predict the classification of a novel instance by determining the majority vote of its nearest neighbors based on the Euclidean distance calculated for each attribute [14]. Nevertheless, the accuracy of the model can be influenced by the choice of distance metric used to calculate nearest neighbors. Figureure (6) shows the prediction of KNN. The number of true predictions is 868 while the number of false predictions is 739. The training accuracy of this model is 57% and the testing accuracy is 54.3%. It has an AUC test value of 58.7%. In addition, a precision of 57%, a recall of 57%, and an F1 score of 56.9%.

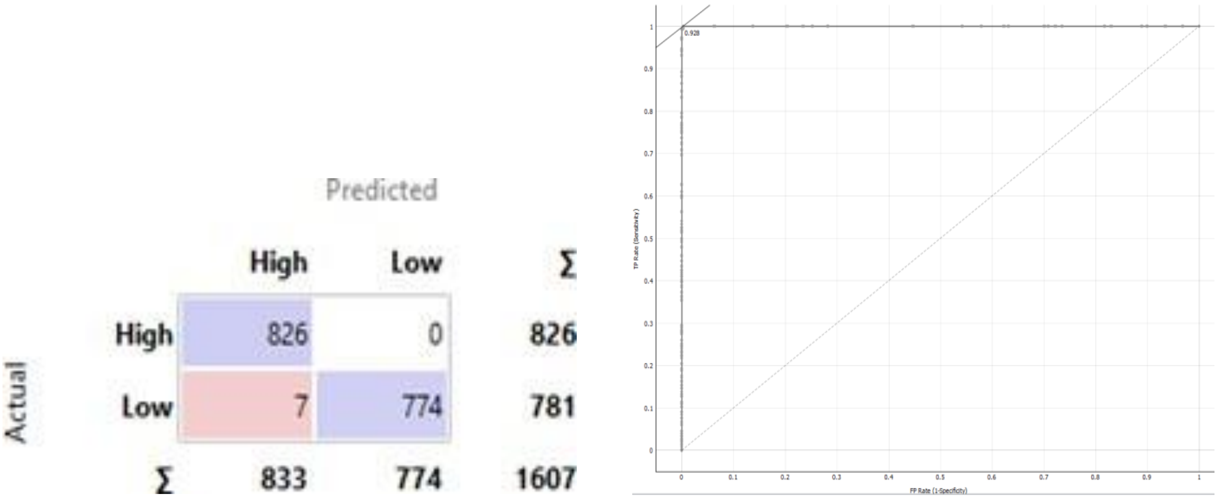


		Predicted		
		High	Low	Σ
Actual	High	455.4	370.6	826
	Low	368.4	412.6	781
	Σ	824	783	1607

Figure.6. Confusion matrix and ROC plot of KNN model.

Cat Boost

CatBoost is a gradient boosting technique for decision trees developed by engineers and researchers at Yandex and is available as an open source tool. Its effective use in a variety of areas, including search, recommendation systems, personal assistants, self-driving vehicles, weather forecasting and other Yandex-related operations, as well as other companies such as CERN, Cloudflare and Careem Taxi, is a testament to its adaptability. The tool is freely accessible to anyone interested. CatBoost has two key advantages: unlike other ML methods, it delivers produces cutting edge results without the need for extensive data training, and it provides excellent out-of-the-box support for many of the more complex data formats commonly found in business environments [15]. Figureure (7) shows the CatBoost classifier. The number of accurate predictions is 1600 while the number of inaccurate predictions is 7, which is the highest value compared to previous models. This model has a training accuracy of 99.6% and a testing accuracy of 97.1%. The AUC score is 100%, the recall is 99.6% and the F1 score is 99.6%.

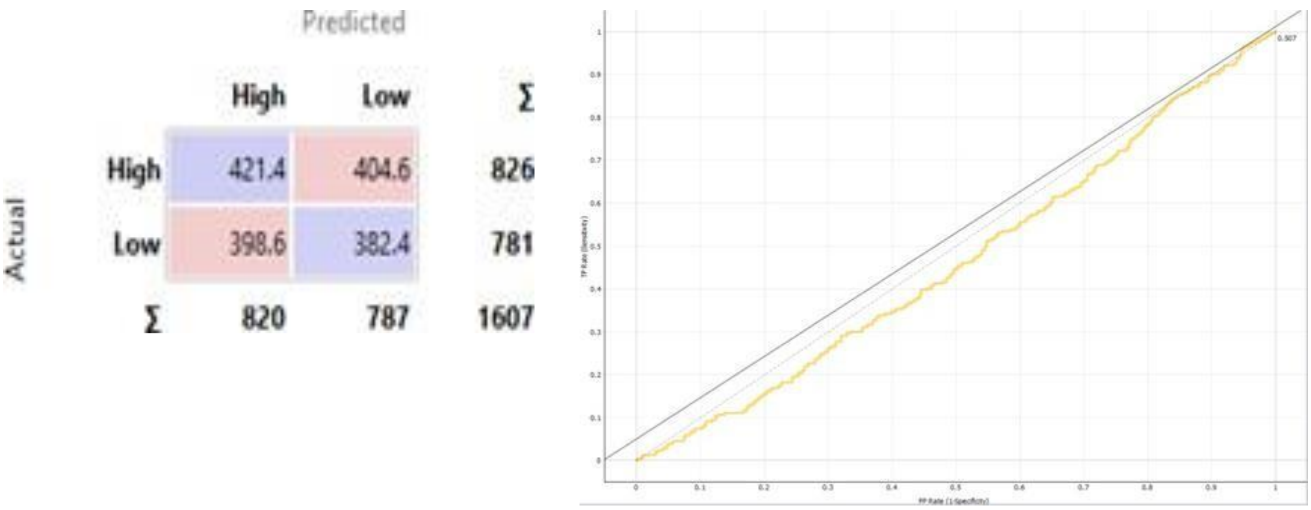


		Predicted		
		High	Low	Σ
Actual	High	826	0	826
	Low	7	774	781
	Σ	833	774	1607

Figure.7. Confusion matrix and ROC plot of CatBoost model.

Logistic Regression (LR)

LR is another well-known probabilistic statistical model for solving classification problems. Probabilities are usually calculated using a logistic function commonly known as a sigmoid function. The LR hypothesis tends to limit the function to values between 0 and 1. (Tolles and Meurer, 2016). This classifier evaluates the relationship between a categorical dependent variable and one or more independent factors in a data set. The dependent variable is the class of target values for the prediction. Figureure (8) shows the prediction of LR. The predicted results are displayed in a confusion matrix and the calculated performance of the model has also been illustrated. The total number of accurate predictions is 804, while the total number of incorrect predictions is 803. The training accuracy is 51.4% and the testing accuracy is 51.2%. Furthermore, it has a test AUC score, precision, recall and F1 score of 46.6%, 26.4%, 51.4% and 34.9%, respectively.

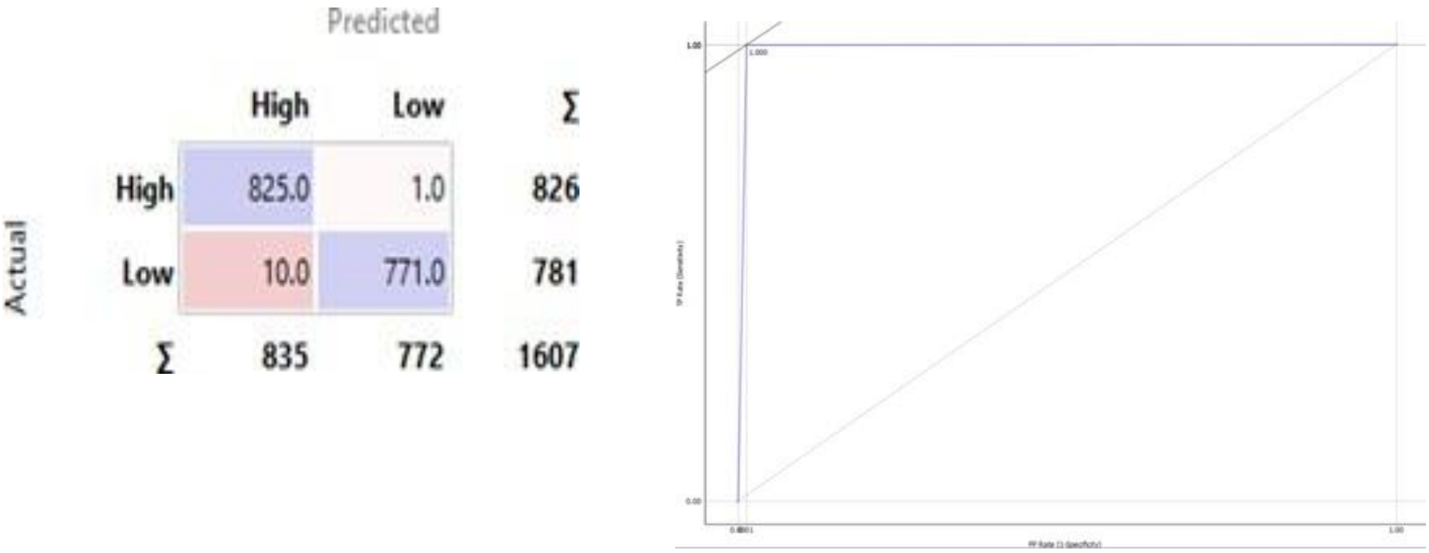


		Predicted		
		High	Low	Σ
Actual	High	421.4	404.6	826
	Low	398.6	382.4	781
	Σ	820	787	1607

Figure.8. Confusion matrix and ROC plot of LR model.

AdaBoost

Adaboost is one of several ensemble-expanding methods. It collects multiple classifiers to improve their accuracy. Figureure (9) shows Ada's prediction. The predicted results are shown in the confusion matrix, where the total number of accurate predictions is 1596 while the number of incorrect predictions is 11. The training accuracy is 99.3% and the testing accuracy is 94.6%. The AUC value of 99.3% and other performance results are consistent with the accuracy value.



		Predicted		
		High	Low	Σ
Actual	High	825.0	1.0	826
	Low	10.0	771.0	781
	Σ	835	772	1607

Figure. 9. Confusion matrix and ROC plot of Ada model.

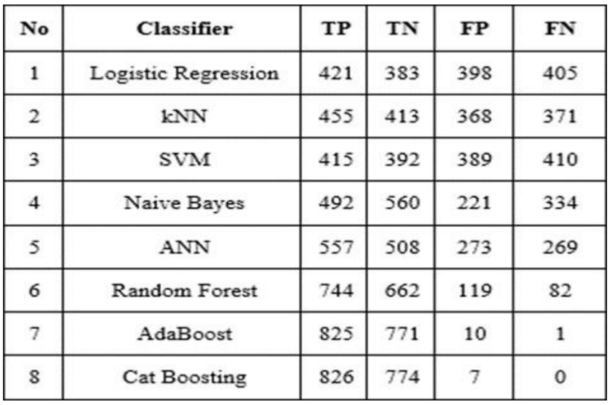
Integrating a machine learning-based model to predict leakage settings requires developing a robust and accurate model, incorporating it into water operations, training non-revenue water teams in its use, reduces the impact on the health of the water drinking network. A study published in Sweden has shown that an ML leak prediction model is highly accurate and can be integrated into the practice of the Non-Revenue Water team to help water utilities identify leaks and high-risk WDNs for rehabilitation plans.

RESULT AND ANALYSIS

This section displays the results of the Naive Bayes, RF, KNN, SVM, ANN, LR, Decision Tree, Adaboost, and CatBoost algorithms. The parameters used to evaluate the analysis of the algorithm were accuracy score, precision, recall and F-1.

The confusion matrix represents a summary of the predictions in matrix form and indicates the number of correct and incorrect predictions for each class. This made it easier to identify classes that were incorrectly classified as another class and the appropriate formulas were used accordingly. The matrix includes the following metrics. TP as a leakage has been detected with a high-risk condition. TN as a negative report No leakages were detected with a High-risk condition. FP as positive report leakages was detected with a Low-risk condition. FN as a negative report that No leakages was detected with a Low-risk condition. As shown in Table (2).

Table 2: Confusion Matrix Results



No	Classifier	TP	TN	FP	FN
1	Logistic Regression	421	383	398	405
2	kNN	455	413	368	371
3	SVM	415	392	389	410
4	Naive Bayes	492	560	221	334
5	ANN	557	508	273	269
6	Random Forest	744	662	119	82
7	AdaBoost	825	771	10	1
8	Cat Boosting	826	774	7	0

The algorithm with the highest accuracy usually produces more accurate results. Table (3) shows the accuracy of each method.

Table 3: Model Accuracy

No	Classifier	Accuracy
1	Logistic Regression	51.40%
2	kNN	57.00%
3	SVM	58.40%
4	Naive Bayes	66.50%
5	ANN	72.30%
6	Random Forest	96.50%
7	AdaBoost	99.30%
8	Cat Boosting	99.60%

The CatBoost algorithm has demonstrated superior performance compared to other algorithms, achieving an accuracy score of 99.6%. The accuracy metric is derived from the features extracted from the datasets, and improvements to both the datasets and computational tools can lead to further improvements in accuracy. The selection of the optimal algorithm for prediction is based on the accuracy values generated by the algorithms. Figureure (10) shows the accuracy graph for eight ML techniques.

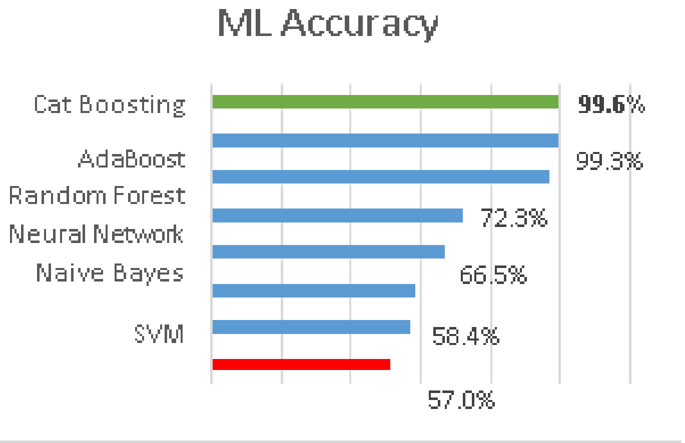
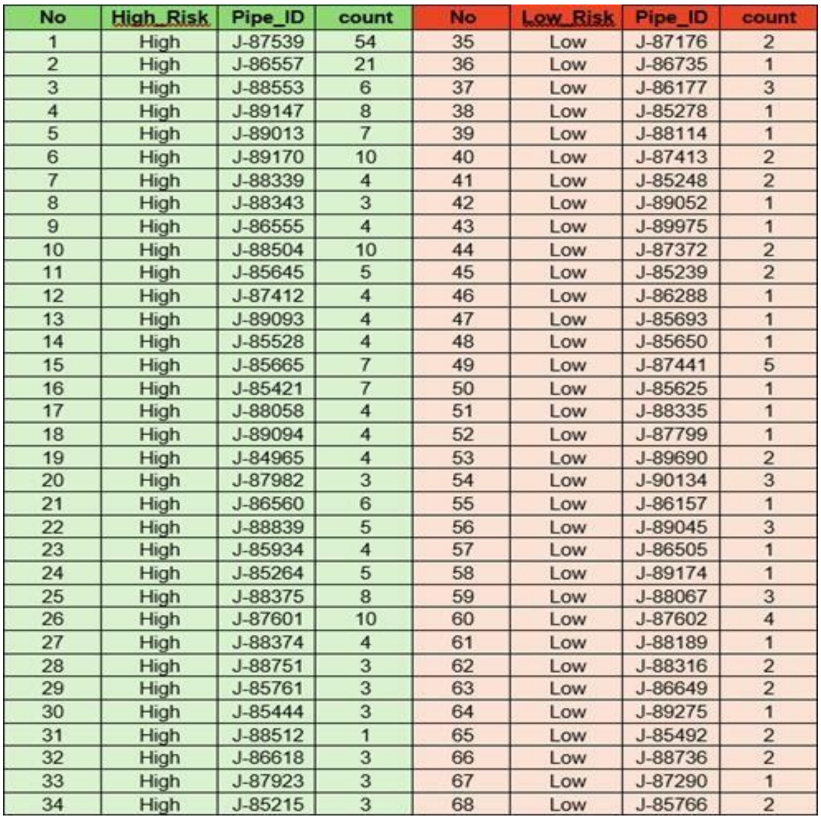


Figure. 10. ML accuracy comparison

PREDICTION MODEL

According to the previous ML Figureure (10), we can see that the highest accuracy is Cat Boost with 99.6% and the lowest is LR with 51.4% accuracy. As a result of the prediction model Table (4), we can see that the high risk has the highest value is J-87539 with count 54 and the lowest is J- 85215. On the other hand, the highest low risk value is J-87176 with count 2 and the lowest is J- 86597.

Table 4: Prediction results via CatBoost classifier.



No	High_Risk	Pipe_ID	count	No	Low_Risk	Pipe_ID	count
1	High	J-87539	54	35	Low	J-87176	2
2	High	J-86557	21	36	Low	J-86735	1
3	High	J-88553	6	37	Low	J-86177	3
4	High	J-89147	8	38	Low	J-85278	1
5	High	J-89013	7	39	Low	J-88114	1
6	High	J-89170	10	40	Low	J-87413	2
7	High	J-88339	4	41	Low	J-85248	2
8	High	J-88343	3	42	Low	J-89052	1
9	High	J-86555	4	43	Low	J-89975	1
10	High	J-88504	10	44	Low	J-87372	2
11	High	J-85645	5	45	Low	J-85239	2
12	High	J-87412	4	46	Low	J-86288	1
13	High	J-89093	4	47	Low	J-85693	1
14	High	J-85528	4	48	Low	J-85650	1
15	High	J-85665	7	49	Low	J-87441	5
16	High	J-85421	7	50	Low	J-85625	1
17	High	J-88058	4	51	Low	J-88335	1
18	High	J-89094	4	52	Low	J-87799	1
19	High	J-84965	4	53	Low	J-89690	2
20	High	J-87982	3	54	Low	J-90134	3
21	High	J-86560	6	55	Low	J-86157	1
22	High	J-88839	5	56	Low	J-89045	3
23	High	J-85934	4	57	Low	J-86505	1
24	High	J-85264	5	58	Low	J-89174	1
25	High	J-88375	8	59	Low	J-88067	3
26	High	J-87601	10	60	Low	J-87602	4
27	High	J-88374	4	61	Low	J-88189	1
28	High	J-88751	3	62	Low	J-88316	2
29	High	J-85761	3	63	Low	J-86649	2
30	High	J-85444	3	64	Low	J-89275	1
31	High	J-88512	1	65	Low	J-85492	2
32	High	J-86618	3	66	Low	J-88736	2
33	High	J-87923	3	67	Low	J-87290	1
34	High	J-85215	3	68	Low	J-85766	2

The rehabilitation strategy should follow a descending order of risk and ensure that resources are allocated to the most vulnerable sections first, thereby significantly reducing the potential for catastrophic failures. Incorporating this targeted approach into this research will ensure that the most critical areas of the network are modernized in a timely manner, thereby improving the overall resilience and efficiency of the water distribution system. As shown in Figureure (11).

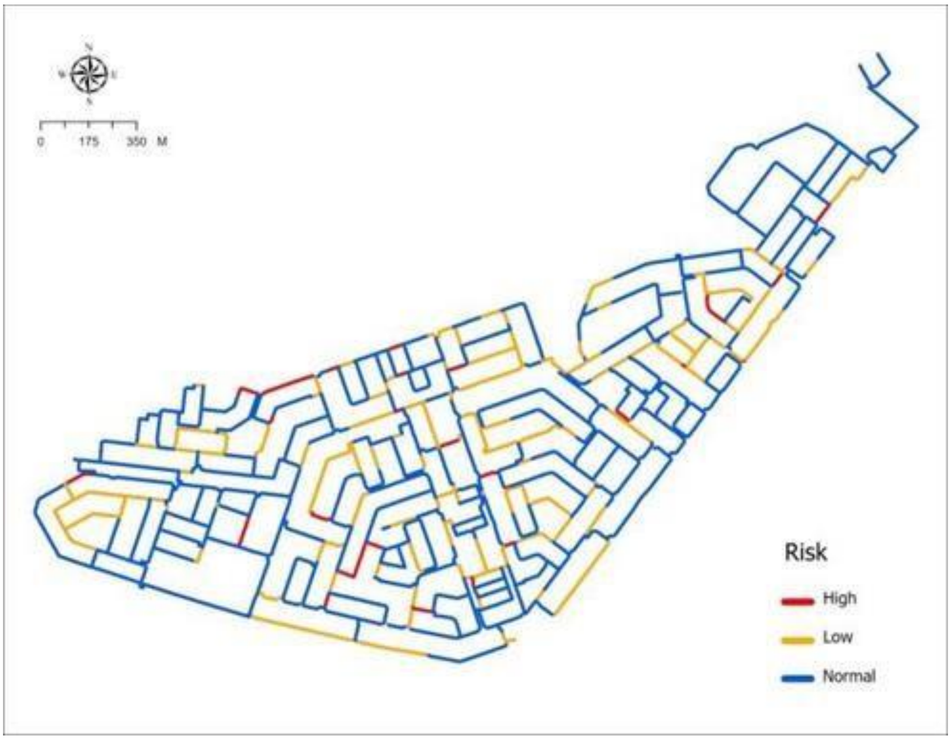


Figure.11. AOI Distribution Network Risk.

CONCLUSION

The critical global problem of water loss in distribution networks proposes a shift from reactive to predictive leak detection using advanced data analytics and machine learning (ML). With a focus on efficiency and sustainability [16], the study analyzes various ML algorithms and identifies as a highlight the 99.6 percent accuracy of the CatBoost classifier in predicting leaks. The resulting predictive model organizes the network into risk-based sections, guiding prioritized remediation and maintenance. This strategy improves the operational process, limits potential water losses, and extends the life of the infrastructure. Ultimately, the approach offers water suppliers the opportunity to reduce water losses non- revenue water (NRW), reduce operating costs and support sustainable water management.

FINDINGS AND LEARNINGS

This study highlights the utility of machine learning in improving water leak detection in water distribution networks (WDNs), with the CatBoost classifier achieving a remarkable accuracy of 99.6%. Data exploration and analysis highlighted the impact of understanding variable interrelationships such as pressure and demand patterns on leak risks and promoted improved data- driven management strategies. The research focuses on strategic resource allocation in high-risk areas to improve water conservation and infrastructure durability. Additionally, it addresses challenges such as model overfitting, which affects generalization to new data, and emphasizes the importance of rigorous model selection and validation. The adaptability and scalability of models are also highlighted as crucial for application in different WDN contexts, supporting the study's alignment with global sustainability goals and reducing environmental impacts in WDN operations.

LIMITATION AND RECOMMENDATION FOR FUTURE RESEARCH

While the present study is groundbreaking in its approach to integrating machine learning for water loss management, it acknowledges certain limitations. Overwhelmingly, it was anchored on a finite data set that, while detailed, may not capture the full variability of global water distribution systems. There was also a risk of overfitting the model, potentially limiting the wider application of the results. Furthermore, the retrospective nature of the analysis may not accurately predict future leaks, particularly as infrastructure and environmental contexts change. Given these limitations, future research avenues emerge. Expanding the datasets to a wider variety of WDNs would improve the robustness and applicability of the models. Leveraging real-time data streams by integrating IoT technologies could improve models' responsiveness to live network changes.

Exploring hybrid machine learning models could also uncover sophisticated patterns in data that surpass the predictive power of singular algorithm frameworks. Ensuring the generalizability of the model remains paramount, requiring rigorous validation processes and research into transfer learning techniques. Furthermore, a comprehensive environmental impact assessment along with a cost-benefit analysis would provide a comprehensive understanding of the economic and environmental aspects of Machine Learning deployment in water distribution networks. As ML solutions eventually carve out a niche in public utilities, the ethical and governance implications must be carefully considered to ensure that such automated systems operate with transparency and accountability and maintain public trust. By considering these aspects, subsequent research can strengthen the link between technology and water resource management and advance the evolution of water systems toward greater resilience and sustainability.

REFERENCES

1. B. K. Mishra, P. Kumar, C. Saraswat, S. Chakraborty, and A. Gautam, ‘Water security in a changing environment: Concept, challenges and solutions’, Water (Switzerland), vol. 13, no. 4, Feb. 2021, doi: 10.3390/w13040490.
2. A. Candelieri, F. Archetti, and E. Messina, ‘Improving leakage management in urban water distribution networks through data analytics and hydraulic simulation’, WIT Transactions on Ecology and the Environment, vol. 171, pp. 107–117, 2013, doi: 10.2495/WRM130101.
3. W. Schultz, S. Javey, and A. Sorokina, ‘Smart Water Meters and Data Analytics Decrease Wasted Water Due to Leaks’, J Am Water Works Assoc, vol. 110, no. 11, pp. E24–E30, Nov. 2018, doi: 10.1002/awwa.1124.
4. P. Głomb, M. Cholewa, W. Koral, A. Madej, and M. Romaszewski, ‘Detection of emergent leaks using machine learning approaches’, Water Supply, vol. 23, no. 6, pp. 2371–2386, Jun. 2023, doi: 10.2166/ws.2023.118.
5. R. Perez et al., ‘Leak localization in water networks: A model-based methodology using pressure sensors applied to a real network in barcelona [applications of control]’, IEEE Control Syst, vol. 34, no. 4, pp. 24–36, 2014, doi: 10.1109/MCS.2014.2320336.
6. Y. Liu, X. Ma, Y. Li, Y. Tie, Y. Zhang, and J. Gao, ‘Water pipeline leakage detection based on machine learning and wireless sensor networks’, Sensors (Switzerland), vol. 19, no. 23, Dec. 2019, doi: 10.3390/s19235086.
7. D. Shravani, P. Y R, P. S B, G. R. Salanke, S. G, and F. S. Ahmed, A Machine Learning Approach to Water Leak Localization. 2019.
8. J. Alves Coelho, A. Glória, and P. Sebastião, ‘Precise Water Leak Detection Using Machine Learning and Real-Time Sensor Data’, Internet of Things, vol. 1, no. 2, pp. 474–493, Dec. 2020, doi: 10.3390/iot1020026.
9. F. Ebisi, I. P. Nikolakakos, J. V. Karunamurthi, A. N. Ahmed Binahmed Alnuaimi, E. Al Buraimi, and S. Alblooshi, ‘Machine Learning Schemes for Leak Detection in IoT-enabled Water Transmission System’, in 2023 International Conference on IT Innovation and Knowledge Discovery, ITIKD 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ITIKD56332.2023.10100175.
10. A. Cutler, D. R. Cutler, and J. R. Stevens, ‘Random Forests’, in Ensemble Machine Learning, New York, NY: Springer New York, 2012, pp. 157–175. doi: 10.1007/978-1-4419-9326-7_5.
11. D. Berrar, ‘Bayes’ theorem and naive bayes classifier’, in Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, vol. 1–3, Elsevier, 2018, pp. 403–412. doi: 10.1016/B978-0-12-809633- 8.20473-1.

12. A. D. Dongare, R. R. Kharde, and A. D. Kachare, ‘Introduction to Artificial Neural Network’, 2012.
13. Y. Xu, E. Xu, and S. Lan, ‘Leakage Diagnosis of Water Supply Network by SVM’, in Proceedings - 2020 International Conference on Artificial Intelligence and Computer Engineering, ICAICE 2020, Institute of Electrical and Electronics Engineers Inc., Oct. 2020, pp. 94–97. doi: 10.1109/ICAICE51518.2020.00024.
14. O. Harrison, ‘Machine Learning Basics with the K- Nearest Neighbors Algorithm’, 2018. [Online]. Available: https://towardsdatascience.com/machine learning-basics- with-the-k-nearest-neighbors-algorithm-6a6e71d01761
15. S. Ben Jabeur, C. Gharib, S. Mefteh-Wali, and W. Ben Arfi, ‘CatBoost model and artificial intelligence techniques for corporate failure prediction’, Technol Forecast Soc Change, vol. 166, May 2021, doi: 10.1016/j.techfore.2021.120658.
16. X. Wan, P. Khorsandi Kuhanestani, R. Farmani, E. Keedwell, and P. D. Student, ‘Literature Review of Data Analytics for Leak Detection in Water Distribution Networks: A Focus on Pressure and Flow Smart Sensors’, 2022.