Anomaly detection for water meter monitoring with machine learning

Yung Hsin, Lin, Faculty of Science, Queensland University of Technology, Brisbane

Abstract

This study developed a hybrid machine learning model for detecting anomalies in water meter data, addressing the issue of water wastage through effective monitoring. The research aims to identify and predict unusual household water consumption patterns by using a combination of unsupervised and supervised learning methods, specifically density-based spatial clustering of applications with noise (DBSCAN) for identifying and long short-term memory (LSTM) for the time-series prediction. The methodology includes preprocessing steps, anomaly feature extraction using DBSCAN, supervised learning with LSTM to predict anomalies and the final evaluation. The dataset contains hourly water meter readings for 107 users over a 6-month period in 2016, obtained from online historical sources. The results show that the model achieves 71% accuracy, capable of capturing many anomalies. Despite low precision and a high false positive rate, the model demonstrates strong sensitivity in detecting potential anomalies. The sensitivity of this model shows the potential of early water anomaly detection and also provides a reference for using hybrid machine learning models for water anomaly detection. It is recommended that future work focuses on integrating additional features, consulting with water management experts, and optimizing model parameters to improve performance and applicability.

Keywords

Anomaly detection, water meter monitoring, machine learning, DBSCAN, LSTM.

1. Introduction

This paper explores innovative machine learning techniques applied in water monitoring management. In recent years, water waste has been a vital environmental issue in resource sustainability. To tackle the issue, it is essential to facilitate water monitoring management by tracking water usage meter profiles from households or industries and identifying anomaly usage. A key problem within water monitoring management is the difficulty in effectively identifying usage patterns and detecting abnormal usage from water meter data, leading to potential water waste. Implementing machine learning techniques can significantly increase the efficiency and accuracy of these monitoring systems and save the costs of manual monitoring methods. However, the absence of advanced AI algorithms integrated into water monitoring management to identify usage patterns, detect anomalous usage, and address maintenance requirements based on water meter data is still a significant challenge to the efficiency of the water monitoring platform. Therefore, the objective of this research is to identify unusual household water consumption patterns on a daily and weekly basis through the implementation of machine learning algorithms. Specifically, the study aims to develop a hybrid model, that combines the advantages of supervised and unsupervised methods, utilizing unsupervised methods to detect anomaly features and develops a supervised model to make

predictions based on time-series meter data.

Previous studies have investigated various applications of machine learning methods for water monitoring management, including supervised learning methods, self-learning methods and unsupervised learning methods, etc. Different from existing knowledge, the novelty of the hybrid model in this paper lies in its combination of supervised and unsupervised learning methods. It improves both accuracy and efficiency compared to solely unsupervised or supervised models. This study addresses two questions based on the research problem: (1) How to build a hybrid model combining both supervised and unsupervised machine learning methods for water anomaly identification based on historical data? (2) What level of performance metrics, including accuracy, precision, recall rate and F1 score, can be achieved by the resulting model for water anomaly identification? The scope of the research focuses on applying machine learning techniques specifically to household water consumption meter data.

The paper went through a comprehensive data analysis process to implement a hybrid machine learning model. The process involves data collection, data preprocessing, the application of DBSCAN and LSTM models and the evaluation. If successful, this research could contribute to future research by demonstrating the feasibility of a hybrid model combining DBSCAN and LSTM for water anomaly detection in water meter monitoring. While previous literatures have explored various machine learning models for water monitoring issues, few studies have developed a combination model that utilizes DBSCAN as the unsupervised learning method to identify the features and LSTMs as the supervised learning method for learning these features to detect anomalies on new data. The combination approach has the potential to provide a stronger and accurate prediction for water monitoring. In addition, this paper implements the model with real-time historical data, adding practical value to the model's applicability in real-world. If the model proves reliable and achieves high performance, it would be a valuable tool for real-world water monitoring management.

The rest of the paper is structured as follows: Literature reviews provides an analysis of using machine learning methods in water monitoring field from previous studies. The methodology section introduces the research process and the selected machine learning methods. In the results and discussion section, the output of the model and the reflection of its performance are presented. In the conclusion section, it summarizes the study's objective, process, outcome, limitations and suggestions for future research.

2.Related Work (Literature Review)

The research topic focuses on detecting anomalies in water meter data using machine learning techniques. Previous research has explored various machine learning techniques to detect anomalous usage in water distribution systems or water consumption monitoring. For the forecasting purpose on time series data, machine learning methods can be involved several aspects: unsupervised learning methods, supervised learning methods, and hybrid methods.

2.1 Supervised learning methods

The most commonly used supervised learning methods in water anomaly detection were neural network-based methods, such as artificial neural network (ANN), long short-term memory (LSTMs), and supervised logistic regression and random forests (RF) (Mashhadi, Shahrour, et al., 2022; Inoue,

Yamagata, et al., 2017). NN-based methods are significant in water demand prediction on the large meter datasets (Rahim, Nguyen, et al., 2020). LSTM is particularly capable of solving complex, long-time lag tasks due to its sequential data processing abilities (Hochreiter, 1997). In water monitoring research, Qian, Jiang, et al. (2020) applied the LSTM model to anomaly detection in the water distribution systems(WDS). They focused on employing both sequence-to-point learning paradigm and data balancing methods on the WDS monitoring data to enhance the performance of LSTM models, achieving an F1 score of 78%.

ANN has been shown to achieve excellent performance for the water leakage localization in the water network (Mashhadi, Shahrour, et al.,2022). Alves Coelho, Glória & Sebastião et al. (2020) compared six different machine learning models to localize water leakage, including logistic regression, decision tree, random forest, hierarchical classification, a combination of the principal component analysis (PCA) with K-means clustering, and ANN. Their findings indicated that random forests achieved the best accuracy in almost every scenario, reaching approximately 85% of accuracy.

2.2 Unsupervised learning methods

For analysis of raw time-series data, instead of using supervised learning algorithms, unsupervised learning is more suitable for handling unlabeled data (Inoue, Yamagata, et al., 2017). For the analysis of usage patterns, unsupervised learning such as one-class SVM and clustering are frequently applied.

According to research from Inoue, Yamagata, et al. (2017), they compared two unsupervised methods for anomaly detection, a deep neural network (DNN) with LSTM and a one-class Support Vector Machine (SVM). Their results revealed that the combination of the DNN model led to fewer false positives compared to one class SVM. Arsene, Predescu, et al. (2022) used K-means clustering for data processing to extract the water consumption patterns followed by a comparison of four classification models for prediction, including decision tree (DT), random forests (RF), multilayer perceptron (MLP), and recurrent neural network (RNN) using LSTM cells. Another clustering method, density-based spatial clustering of applications with noise (DBSCAN) was conducted by Ghamkhar, Ghazizadeh, et al. (2023) and successfully captured 98% of abnormal users in the time series dataset.

In conclusion, NN-based model performed well on complex data in previous papers, demonstrating high performance in predicting the anomaly water consumption. Unsupervised learning methods, including One-Class SVM and clustering, were conducted in some previous studies to categorize the unlabelled water dataset effectively. However, even with the popular machine learning algorithms, results may vary significantly based on the quality and the type of the dataset as well as the purpose of the analytics. A significant gap in prior studies is that most of the papers used the simulated data to detect irregular water usage rather than using historical data. This limitation can lead to low accuracy of models and their applicability in the real-world scenarios.

Given the advantages and disadvantages of the above machine learning methods, as well as the gap identified in previous research, this research proposes a hybrid model, using DBSCAN as the unsupervised method to identify the patterns and data labelling, followed by LSTM as supervised method to conduct prediction, in order to enhance the model prediction ability. Specifically, the clustering method, DBSCAN, is used first to extract abnormal usages, making good use of its efficiency in handling time complexity, while LSTM is then employed as supervised learning

method to perform forecasting based on the extracted features for its ability of processing time series sequential data. Additionally, in order to enhance the accuracy of detection on unusual water usage, this research is using historical time-series data rather than simulated data to develop machine learning models. The novelty of this study is to implement a hybrid model that combines DBSCAN and LSTM for anomaly detection with time series historical data. This approach collects the advantages of supervised learning and unsupervised learning to increase the accuracy of anomaly identification and provide as a practical insight that is applicable in the real-world water monitoring management, contributing valuable knowledge to this domain.

3. Methods

Based on the results in previous studies, this research aims to develop a hybrid model combining DBSCAN and LSTM to detect anomalies in water meter data. DBSCAN is selected for its ability to identify noise effectively, making it ideal for the unsupervised machine learning component. LSTM, a neural network, is used for handling complex time-series data.

The research objective is to build a hybrid model which combines the advantages of supervised and unsupervised methods to effectively detect water anomaly in time-series meter data. The hybrid model is designed to employ clustering using DBSCAN to identify anomaly features and then build a LSTM model to conduct predict unusual water usage. This study adopts an artifact-oriented approach, demonstrating the data analytic process which includes data collection, first data preprocessing, unsupervised method (DBSCAN), second data preprocessing, supervised method (LSTM), and evaluation. The implementation tool is the Python programming language on Google Colab.

3.1 Data Collection

Due to privacy concerns, most household water meter data is not publicly available. However, to ensure real-world applicability of the results, it is preferable to use historical time series data over simulated data. Accordingly, this research utilizes data from the HELIX (Hellenic Data Service) open data online page, specifically from the FP7 project DAIAD. This dataset collected from Smart Water Meter Consumption time-series data for 1,007 randomly selected consumers from approximately 110,000 available smart water meters. The original dataset spans from 01/01/2015 at 00:00 to 19/05/2017 at 23:59, containing hourly water consumption measurement in households. Due to memory limitations and high dimensional data, six months of 2016 data were extracted, including 10% of consumers, resulting in 361,712 records.

The data includes four attributes: user key, datetime, meter reading, and diff. There are 107 users with unique user keys in the filtered dataset. The datetime attribute contains hourly data from January to June 2016. The meter reading indicates the total volume of consumed water, and the diff represents the water consumption difference from the previous measurement.

3.2 First Data Preprocessing

The first preprocessing mainly focused on data cleaning and preparation, which involves dropping missing and duplicate values, changing data types, and performing feature engineering by categorizing time periods (e.g., morning, afternoon, night, and days of the week). The dataset was also standardized to improve the model's results.

3.3 Unsupervised Method (DBSCAN)

According to the literature review (Ghamkhar, Ghazizadeh, et al., 2023; Ester, Kriegel et al., 1996), DBSCAN excels at identifying noise by assigning data points to different clusters based on their similarity. Unlike other clustering methods, such as K-means, DBSCAN groups the data and determines the number of clusters itself without specifying the number of clusters first. It also classifies points which are not assigned to any cluster as a noise class. Since the number of clusters is not required to specify in advance, DBSCAN is very sensitive to hyperparameter settings (eps and min_samples). According to the definition from Ghamkhar, Ghazizadeh, et al. (2023), Epsilon (eps) defines the maximum distance between two samples within the same neighborhood. The minimum number of samples (min_samples) indicates the number of samples required in the neighborhood of a core point to determine cluster density.

3.4 Second Data Preprocessing

After DBSCAN clustering, further preprocessing was performed. In some clusters, there are some outliers that can be seen as anomalies because they have low similarity within the same cluster. Therefore, the outliers within each cluster were identified using the Interquartile Range (IQR) method. The definition of outliers using IQR by OpenAI (2023):

Q1 = 0.25 quantile

Q3 = 0.75 quantile

IQR= Q3 − Q1

Lower Bound = Q1− 1.5 × IQR

Upper Bound = Q3 + 1.5 × IQR

Outliers = { x < Lower Bound or x > Upper Bound}

These outliers, along with the noise cluster (-1) from DBSCAN, were labeled as anomalies. After labeling, the dataset was split into training (80%) and testing (20%) sets for the next supervised learning training, and the class weights adjustment method, SMOTE in imblearn was applied to address the data imbalance.

3.5 Supervised Method (LSTM)

After the labeled data was prepared, a supervised learning prediction model- LSTMs was then trained to learn features of normal and anomalous readings and to detect abnormal usage on test datasets. LSTM is neural networks-based method, which introduces a memory cell with forget, update and output gates to control the information (Qian, Jiang et al., 2020). The LSTM model used in this paper consisted of a single LSTM layer with 50 hidden units and ReLU activation, followed by a connected dense layer with a sigmoid activation function. The model applied the Adam optimizer with a learning rate of 0.0001 and employs early stopping with a patience of 5 to prevent overfitting. The output was a probability for each test data point indicating the likelihood of it being an anomaly. For example, assuming a new reading (u12345, 31/4/2017, 1450, 3) is input, the model might return a probability of 0.7145, indicating the possibility of anomaly usage.

3.6 Evaluation

To evaluate the results, a confusion matrix and performance metrics are given. Since the prediction results are provided as a possibility number, a threshold must be chosen to distinguish between normal and anomalous data. Predictions with a probability greater than the threshold are classified as anomalies. The confusion matrix and performance metrics, including accuracy, precision, recall rate and F1 score, were then presented based on the threshold. The definition of performance metrics by OpenAI (2023):

Accuracy: This is the overall accuracy, the proportion of samples predicted correctly to all samples.

$$Accuracy = \frac{tp+tn}{tp+fp+fn+tn} \quad (1)$$

Precision: Measures the accuracy of the model in detecting anomalies, reflecting how many of the data predicted as anomalies are actual anomalies.

$$Prscision = \frac{tp}{tp+fn} \quad (2)$$

Recall: Measures correct detections in actual anomaly data, reflecting the situation of missed detections.

$$Recall = \frac{tp}{tp+fp} \quad (3)$$

F1 Score: Balances Precision and Recall and can comprehensively evaluate the overall performance of the model in anomaly detection.

$$F1\ score = \frac{2tp}{2tp+fp+fn} \quad (4)$$

4. Results and Discussion

This section presents the outcomes of the hybrid model according to the methodology above.

4.1 Results

After the first data preprocessing, the dataset consisted of 361,712 readings across 13 features including 'user key', 'datetime', 'meter reading', 'diff', and time-related variables such as parts of the day and days of the week.

4.1.1 DBSCAN Results

Effectively identifying normal and abnormal values is a key factor for prediction results. Therefore, the hyperparameter setting of DBSCAN is essential for accurately labelling data points. The comparison table below shows different hyperparameter combinations and their clustering results.

	% of consumers	eps	min_samples	number of clusters	number of anomaly
6 months	10%	0.1	5	451	9015
	10%	0.3	5	99	1272
	10%	0.5	5	75	658
	10%	0.27	26	55	8267
	10%	0.1	26	43	26731
	10%	0.5	26	57	1883
	10%	0.5	240	29	5480

Table. 1. Different hyperparameter setting in DBSCAN

After experimenting with various hyperparameter combinations, a minimum samples value (min_samples) of 26 was chosen - twice the number of features, as suggested by Sander, Ester, et al. (1998). The eps parameter was set at 0.1, based on the recommendation by Ghamkhar, Ghazizadeh, et al. (2023) that smaller eps results in more noise points being identified. With the combination of min_samples=26 and eps=0.1, the DBSCAN model resulted in 43 clusters, with 26,731 points labeled as anomalies (cluster = -1).

The second data preprocessing focuses mainly on identifying outliers within these clusters, data points labeling and dataset splitting. This phase resulted in 32,136 of anomaly were labeled, accounting for 9% in the entire dataset.

4.1.2 LSTM Results

The LSTM training process is visualized in the loss and accuracy plots. According to the plots, it is obvious that the validation loss is consistently higher than the training loss, while the validation accuracy remains lower than the training loss throughout the epochs. It indicates that the model was overfitting, it learned well on training data but does not perform well on unseen data. Specifically, the model cannot effectively distinguish between normal and abnormal data, which is a problem for anomaly detection tasks.

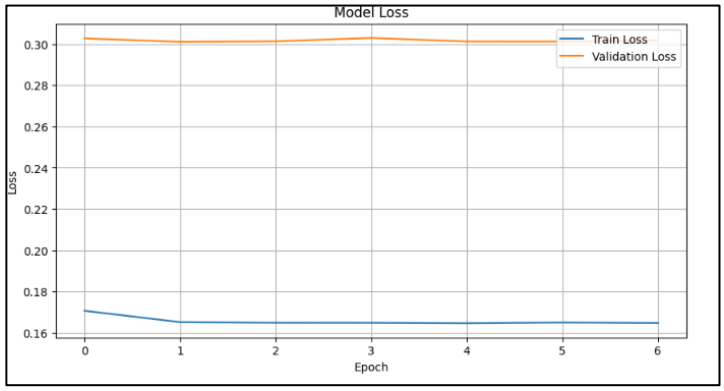

Fig. 1. Loss plot of training and validation

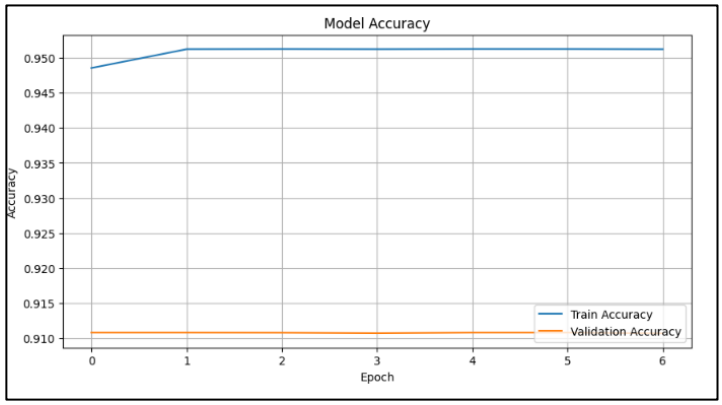

Fig. 2. Accuracy plot of training and validation

4.1.3 Evaluation Results

In the final evaluation step, the results were presented using a confusion matrix and performance metrics. As mentioned in methodology, choosing an appropriate threshold was important to distinguish the normal and abnormal data. After experimenting with different threshold settings, a threshold of 0.092 was selected as it reaches a precision (29%), recall rate (24%) and F1 score (13%), while maintaining an accuracy of 71%.

```
Threshold: 0.092
Accuracy: 0.7193406988481589
Precision: 0.09311621891717124
Recall: 0.24542919119925627
F1 Score: 0.13500958874920094
```

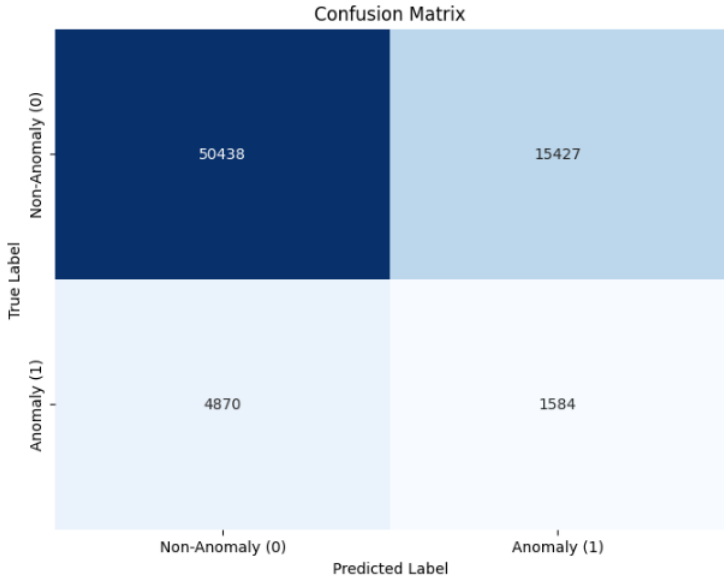

Fig. 3. Confusion Matrix and Performance Metrics

The confusion matrix reveals that most of the non-anomaly usages were correctly identified (True Negative), and 1584 instances were correctly predicted as anomalies (True Positive). However, high false positives and false negatives indicate a high incorrect prediction, in which 15427 normal readings were misrecognized as anomalies and 4870 anomalies were not identified.

Based on the performance metrics and the confusion matrix, the higher recall rate compared with precision suggests that the model predicted many anomaly readings. However, due to the low precision, the F1 score might not be high. In addition, low precision and high FP indicate that many normal usages were incorrectly detected as anomalies.

4.2 Discussion

Overall performance is suboptimal, primarily due to a high false positive rate, which causes low precision, and it negatively impacts the F1 score. The model detected most water anomalies but also misrecognized a lot of normal usage as anomalies, which may cause many normal water users to receive notifications of abnormal water use. However, due to the objective of identifying and predicting anomalous water usage, the main focus was on reducing the misrecognition of true anomaly data and increasing the number of true anomaly data that are recognized. In other words, the purpose of this research is to reduce false negatives (lower-left corner of the confusion matrix) and increase true positives (lower-right corner) with reasonable accuracy. Therefore, the higher recall rate allows for higher sensitivity to anomalies, which could be useful in alerting users to potential water waste.

Different from the other classification prediction models, LSTM provides a proportion of the anomaly class instead of a clear binary classification. Therefore, it brings the flexibility of the model in this research. Stakeholders can adjust the threshold based on their tolerance for false positives. For instance, households may tolerate higher false positives, but industries may want stricter thresholds to avoid disruptions. Another benefit is that the sensitive alert signals from the high false positive can prevent water wastage by allowing users to check for potential leaks or other issues sooner.

4.2.1 Relation to The Research Questions and Literature Review

In relation to the research questions and literature review, this study developed a hybrid model of DBSCAN and LSTM that achieved 71% accuracy, 24% recall, and a 13% F1 score, addressing the research questions. The literature suggests that DBSCAN is effective in capturing usage patterns but requires proper tuning, which was reflected during the research process. Furthermore, LSTM models benefitted time series data, supported by literature indicating that neural networks perform better with complex datasets.

4.2.2 Limitations and Future work

The overfitting and misclassification issues suggest that the DBSCAN model failed to correctly identify unusual cases, which resulted in the LSTM model not effectively learning the patterns. These may be due to data quality issues, such as data imbalance and insufficient features, preventing the model from effectively learning patterns. Extreme data imbalance may lead to unreliable accuracy, as accuracy tends to favour the majority class. Additionally, the lack of features in the dataset does not provide the model with enough information to identify or predict patterns. Another limitation is the lack of water domain knowledge during model development. Water meter data relies on experienced experts to define anomaly ranges and understand local water usage habits. Although

AI algorithms can identify the usage patterns in the unlabelled data, the output still needs to be validated and reviewed by water monitoring managers. Potential solutions may involve consulting water experts to properly label the data, rather than relying solely on DBSCAN and outlier detection to identify anomalous data. Moreover, incorporating additional features such as geographical relationships, weather, and family size into the dataset can also enhance the learning ability of the model. Further fine-tuning in DBSCAN and LSTM parameters may also improve the model performance. The setting of eps parameter is challenging because it requires a deep understanding of the data type. The layer in LSTM can also be defined appropriately to avoid overfitting. Ensuring better data quality and correct labels can significantly improve the model's prediction capabilities and precision.

5. Conclusions

This study aims to address the efficient water anomaly detection problem with machine learning techniques. A hybrid model has been developed that integrates DBSCAN with LSTM to enhance the prediction in water meter data, and leverage the strengths of both unsupervised and supervised learning methods to more accurately identify unusual water usage patterns that may indicate leaks or other water waste issues.

According to the research output, the high recall rate became evident that the model was effective in capturing many anomaly patterns, there were significant challenges related to the quality of the data used. However, the model suffered from a high rate of false positives and low precision, indicating the model consistently misclassified normal behaviors as anomalies. These make the results suboptimal, but the sensitivity of the model may benefit the prevention of water leakage. Despite the result not being ideal, the output not only can help capture potential water unusual patterns in the water monitoring management but also contribute an example of the hybrid machine learning model which combined DBSCAN and LSTM in the water monitoring related research domain. Current limitations include data imbalance and the absence of enriched features in the datasets, which can be solved in the future.

Future research should focus on three main areas: Firstly, combining datasets with additional features such as geographic data, weather conditions, and demographic information may address the limitations of current models. Secondly, working with experts in water monitoring management field may lead to more accurate anomaly labeling and richer interpretations of data types. Finally, refining the parameters of DBSCAN and LSTM through more adjustments will help reduce overfitting and improve the reliability of the model.

In summary, this study provides an example of using machine learning to improve anomaly detection in water meter systems. By addressing the challenges and implementing AI algorithms, there is great potential for the improvement in efficiency of water management systems, which benefiting the goals of sustainability and efficiency in resource management.

Acknowledgements

I would like to express my gratitude to Dr. Wenzong Gao for guilding this research. His suggestion and encouragement is invaluable in this research.

References

[1] Arsene D, Predescu A, Pahonțu B, Chiru CG, Apostol E-S, Truică C-O. Advanced Strategies for Monitoring Water Consumption Patterns in Households Based on IoT and Machine Learning. Water. , vol. 14, no. 14, pp. 2187, 2022. https://doi.org/10.3390/w14142187

[2] J. Inoue, Y. Yamagata, Y. Chen, C. M. Poskitt, and J. Sun, “Anomaly detection for a water treatment system using unsupervised machine learning,” 2017 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 1058-1065, 2017. https://doi.org/10.1109/ICDMW.2017.149.

[3] J. Alves Coelho, A. Glória, and P. Sebastião, “Precise water leak detection using machine learning and real-time sensor data,” IoT, vol. 1, no. 2, pp. 474-493, 2020. https://doi.org/10.3390/iot1020026.

[4] N. Mashhadi, I. Shahrour, N. Attoue, J. El Khattabi, and A. Aljer, “Use of machine learning for leak detection and localization in water distribution systems,” Smart Cities, vol. 4, no. 4, pp. 1293-1315, 2021. https://doi.org/10.3390/smartcities4040069.

[5] R. Magini, M. Moretti, M. A. Boniforti, and R. Guercio, “A machine-learning approach for monitoring water distribution networks (WDNs),” Sustainability, vol. 15, no. 4, 2981, 2023. https://doi.org/10.3390/su15042981.

[6] Z. Y. Wu, A. Chew, X. Meng, J. Cai, J. Pok, R. Kalfarisi, K. C. Lai, S. F. Hew, and J. J. Wong, “Data-driven and model-based framework for smart water grid anomaly detection and localization,” AQUA - Water Infrastructure, Ecosystems and Society, vol. 71, no. 1, pp. 31-41, 2022. https://doi.org/10.2166/aqua.2021.091.

[7] K. Qian, J. Jiang, Y. Ding, and S. Yang, “Deep learning-based anomaly detection in water distribution systems,” 2020 IEEE International Conference on Networking, Sensing and Control (ICNSC), pp. 1-6, 2020. https://doi.org/10.1109/ICNSC48988.2020.9238099.

[8] A. Oluyomi, S. Abedzadeh, S. Bhattacharjee, and S. K. Das, “Unsafe events detection in smart water meter infrastructure via noise-resilient learning,” 2024 ACM/IEEE 15th International Conference on Cyber-Physical Systems (ICCPS), pp. 259-270, 2024. https://doi.org/10.1109/ICCPS61052.2024.00030.

[9] M. S. Rahim, K. A. Nguyen, R. A. Stewart, D. Giurco, and M. Blumenstein, “Machine learning and data analytic techniques in digital water metering: A review,” Water, vol. 12, no. 1, 294, 2020. https://doi.org/10.3390/w12010294.

[10] S. Hochreiter, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735.

[11] H. Ghamkhar, M. J. Ghazizadeh, S. H. Mohajeri, I. Moslehi, and E. Yousefi-Khoshqalb, “An unsupervised method to exploit low-resolution water meter data for detecting end-users with abnormal consumption: Employing the DBSCAN and time series complexity,” Sustainable Cities and Society, vol. 94, 104516, 2023. https://doi.org/10.1016/j.scs.2023.104516.

[12] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD), pp. 226-231, 1996.

[13] J. Sander, M. Ester, H. P. Kriegel, and X. Xu, “Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications,” Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 169-194, 1998.

[14] S. Nofal, A. Alfarrarjeh, and A. Abu Jabal, “A use case of anomaly detection for identifying unusual water consumption in Jordan,” Water Supply, vol. 22, no. 1, pp. 56-64, 2022. https://doi.org/10.2166/ws.2021.135.

[15] J. Merta and J. Fikejz, “Utilization of machine learning to detect sudden water leakage for smart water meter,” 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), pp. 1-6, 2019. https://doi.org/10.1109/RADIOELEK.2019.8733397.

[16] Hellenic Data Service, “Smart water meter consumption time series - datasets,” HELIX, 2020. [Online]. Available: https://data.hellenicdataservice.gr/dataset/78776f38-a58b-4a2a-a8f9-85b964fe5c95.

[17] R. Vanijjirattikhan, S. Khomsay, N. Kitbutrawat, K. Khomsay, U. Supakchukul, S. Udomsuk, and K. Anusart, “AI-based acoustic leak detection in water distribution systems,” Results in Engineering, vol. 15, 100557, 2022.

[18] A. H. Ayati, A. Haghighi, and H. R. Ghafouri, “Machine learning–assisted model for leak detection in water distribution networks using hydraulic transient flows,” Journal of Water Resources Planning and Management, vol. 148, no. 2, 04021104, 2022.

[19] Yang, L., Driscol, J., Sarigai, S., Wu, Q., Lippitt, C. D., and Morgan, M., “Towards synoptic water monitoring systems: A review of AI methods for automating water body detection and water quality monitoring using remote sensing,” Sensors, vol. 22, no. 6, 2416, 2022. https://doi.org/10.3390/s22062416.

[20] OpenAI's ChatGPT, “Discussion on climate change implications.” OpenAI, 2023.