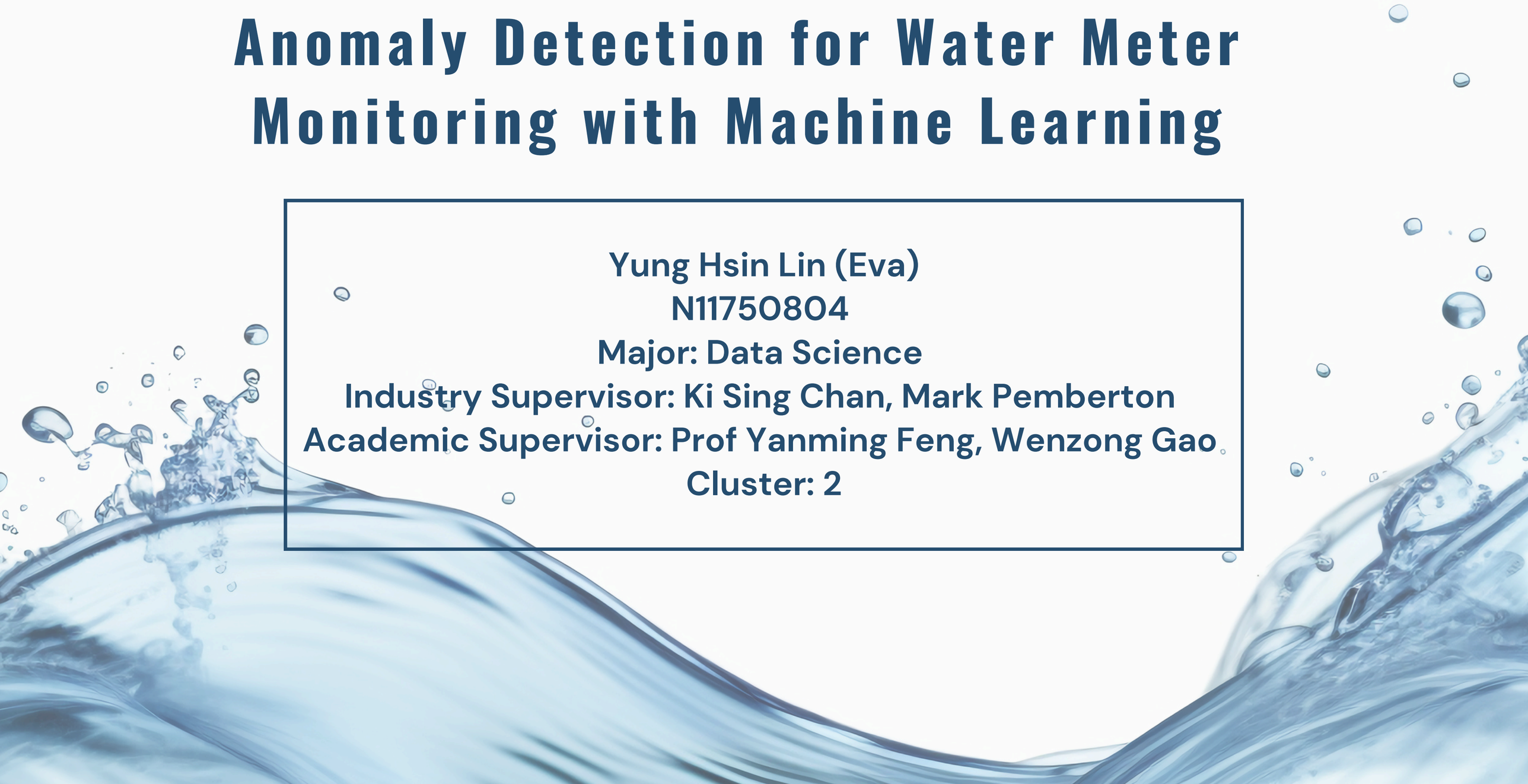
Anomaly Detection for Water Meter Monitoring with Machine Learning

Yung Hsin Lin (Eva)
N11750804
Major: Data Science
Industry Supervisor: Ki Sing Chan, Mark Pemberton
Academic Supervisor: Prof Yanming Feng, Wenzong Gao
Cluster: 2



CONTENT

INTRODUCTION

Background

Water wasting is a vital environmental issue these years.

Identifying unusual usage in households and industries can be helpful in preventing water waste.

Problem

The challenge in water monitoring management is the inability to effectively identify usage patterns and detect abnormal usage from water meter data, leading to potential water waste.

To solve this problem, implementing machine learning techniques can be helpful to enhance efficiency and accuracy, and save manual monitoring methods' costs. However, the lack of efficient machine learning algorithms is still a problem today.

Objective

The research aim is to effectively identify and predict unusual household water usage patterns on a daily and weekly basis using machine learning algorithms.

Different from existing knowledge, the novelty of the hybrid model in this paper is its ability to combine supervised and unsupervised learning methods. It improves both accuracy and efficiency compared to solely unsupervised or supervised models.

LITERATURE REVIEW

Title	Main Methods/Tools	Main Findings	Limitations	Future Suggestions
Anomaly Detection for a Water Treatment System Using Unsupervised Machine Learning	DNN+LSTM, SVM	DNN performed better and had fewer false positives	Limited to specific dataset (SWaT), may not generalize	Use real-time datasets and improve encoding of long-term trends
An unsupervised method to exploit low-resolution water meter data for detecting end-users with abnormal consumption: Employing the DBSCAN and time series complexity	DBSCAN with time series complexity features like LZC, TSLF	Detected 98% of abnormal users, applicable to any dataset	DBSCAN tuning required refinement, process requires expert knowledge	Refine hyperparameters and streamline anomaly detection
Advanced Strategies for Monitoring Water Consumption Patterns in Households Based on IoT and Machine Learning	K-means, Decision tree (DT), Random forests (RF), Multilayer perceptron (MLP) and recurrent neural network (RNN)	Three clusters (sink, shower, toilet), deep learning models performed better	Hot/cold tap separation reduced accuracy, high data volume impacted performance	Expand study to more households, location-based clustering, develop decision support system

OBJECTIVE OF THE RESEARCH

Objective

Building a hybrid model, which combines the advantages of supervised and unsupervised methods, using unsupervised learning to identify leakage features and building a supervised model to conduct prediction on time-series meter data.

Research Question 1

How to build a hybrid model combining both supervised and unsupervised ML methods for water anomaly identification based on historical data?

Research Question 2

What level of performance metrics can be achieved by the resulting model for water anomaly identification?

METHODOLOGY

01

DATA COLLECTION

- Extracted six months of 2016 was collected from HELIX (open-source)
- Sample included 10% of consumers (around 100 users).
- 361,712 readings of hourly household water consumption.

02

1ST DATA PREPROCESSING

- Dropping missing and duplicate values
- Changing data type
- Feature engineering: separates time series into time durations (morning, afternoon, and night), and days of the week.
- Standardization

03

UNSUPERVISED METHOD (DBSCAN)

DBSCAN helps identify unusual patterns in water usage data by assign data to different clusters based on their similarity and find out the noise in the data.

04

2ND DATA PREPROCESSING

- Identified outliers to distinguish between normal and anomalous behaviors within the same cluster.
- Manually label the outliers and class = -1 (noise class classified by DBSCAN) as Anomaly =1
- Split data into training (80%) and test (20%) sets, adjusting class weights for data imbalance.

05

SUPERVISED METHOD (LSTM)

LSTMs is used for analyzing time-series data over time, learning features of normal and anomalous readings and provide a likelihood of abnormal usage on the test set.

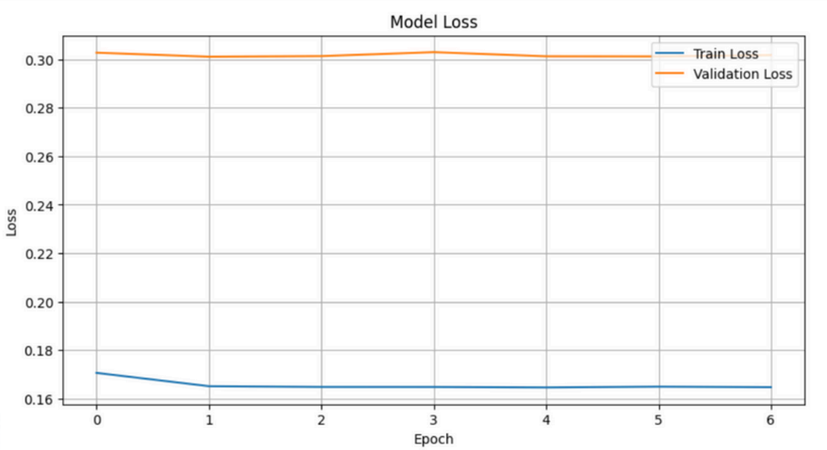
06

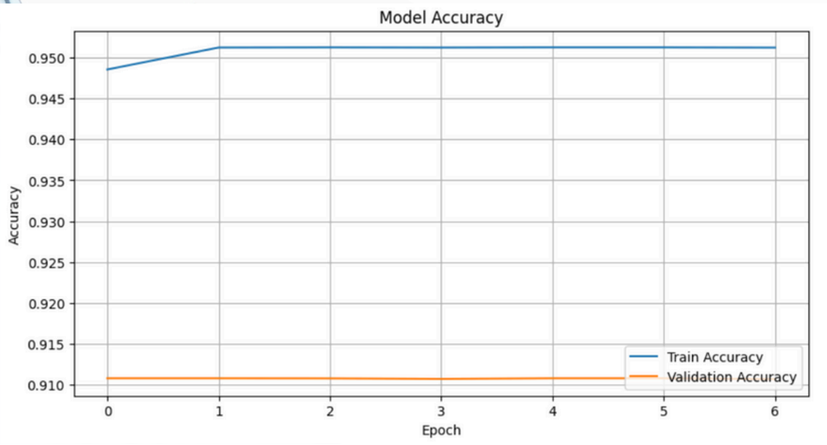
EVALUATION

- A threshold was chosen to distinguish between normal and anomalous data.
- Given a confusion matrix and performance metrics (F1 score, recall rate, precision, and accuracy) based on the selected threshold.

RESULTS

Loss plot and Accuracy plot

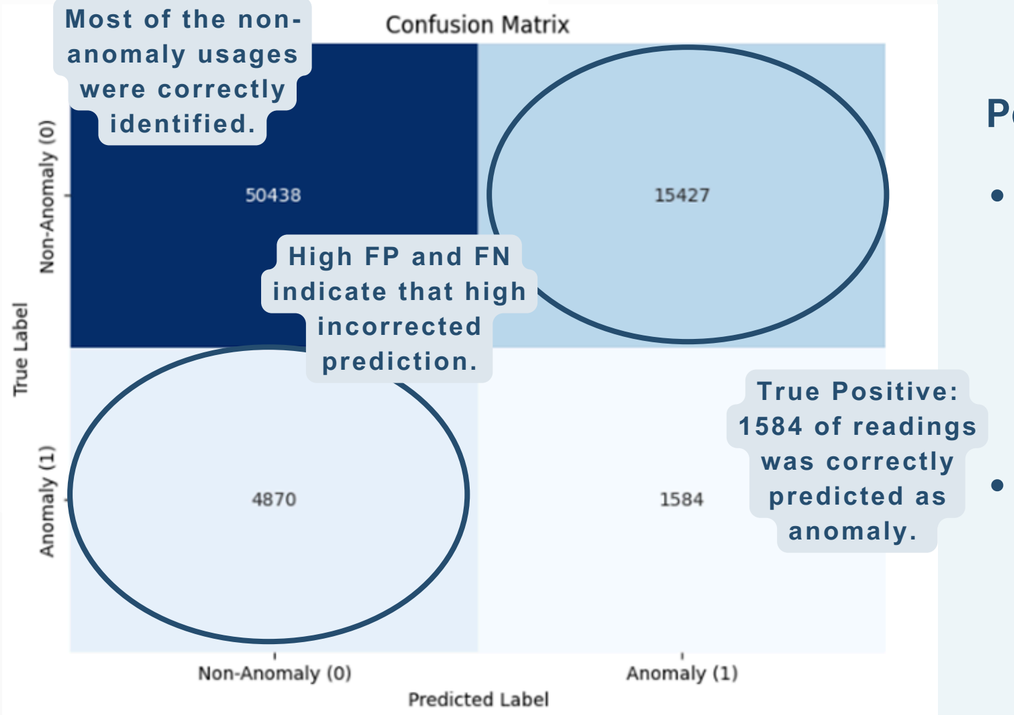




Performance metrics

Predictions with a probability greater than 0.092 are classified as anomalies.

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



LSTM training results:

- Validation loss exceeds training loss.
- Validation accuracy is lower than training accuracy.
- Model is overfitted, it could not learn the pattens effectively.

Performance metrics

- The higher recall rate suggests that the model predicted many anomaly readings. However, due to the low precision, the F1 score might not be high.
- Low precision and high FP indicate that many normal usages were incorrectly identified as anomalies.

DISCUSSION

- Overall performance is suboptimal. The high false positive causes low precision. The model detected most water anomalies but also misrecognized a lot of normal usage as anomalies.

- 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. Therefore, the higher recall rate allows for higher sensitivity to anomalies, which could be useful in alerting users to potential water usage issues.

- The overfitting and incorrect classify problem indicate that the DBSCAN model did not identify the unusual cases correctly which led to the LSTMs model not learning the patterns effectively.

- **Relation to the research questions and literature review:** The study developed a hybrid DBSCAN + LSTM model with 71% precision, 24% recall, and 13% F1 score, replying well to the research questions. Literature indicated DBSCAN is effective in capturing usage patterns but requires tuning, which is reflected in my research as well. LSTM is useful for time series data, which is also supported by literature that neural networks is better in complex datasets.

CONCLUSION

This model provides a predictive function that assigns the possibility of water anomaly events.

Benefits:

- Flexibility: 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.
- Cost savings: 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.

Limitation:

- Lower F1 score suggests future improvements in feature selection methods and model refinement, such as fine-tuning parameters and balancing the dataset
- Consultation with water experts is recommended for accurate data labeling instead of relying solely on DBSCAN and outlier detection.

HIGHLIGHTS OF RESEARCH OUTPUTS

- Overall model performance is suboptimal due to a high false positive rate.
- High recall suggests the model was effective in capturing many anomaly patterns.
- Further tuning is required, particularly in the label identification step with DBSCAN.
- Improvements in precision and overall accuracy are necessary for more reliable predictions.

REFERENCES

- Arsene, D., Predescu, A., Pahonțu, B., Chiru, C. G., Apostol, E.-S., & Truică, C.-O. (2022). Advanced strategies for monitoring water consumption patterns in households based on IoT and machine learning. Water, 14(14), 2187. https://doi.org/10.3390/w14142187
- Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C. M., & Sun, J. (2017). Anomaly detection for a water treatment system using unsupervised machine learning. In 2017 IEEE International Conference on Data Mining Workshops (ICDMW).
- Ghamkhar, H., Jalili Ghazizadeh, M., Mohajeri, S. H., Moslehi, I., & Yousefi-Khoshqalb, E. (2023). 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, 94(1).
- Coelho, J. A., Glória, A., & Sebastião, P. (2020). Precise water leak detection using machine learning and real-time sensor data. IoT, 1(2).
- Mashhadi, N., Shahrour, I., Attoue, N., El Khattabi, J., & Aljer, A. (2021). Use of machine learning for leak detection and localization in water distribution systems. Smart Cities, 4(4).
- Wu, Z. Y., Chew, A., Meng, X., Cai, J., Pok, J., Kalfarisi, R., Lai, K. C., Hew, S. F., & Wong, J. J. (2022). Data-driven and model-based framework for smart water grid anomaly detection and localization. AQUA — Water Infrastructure, Ecosystems and Society, 71(1).
- Rahim, M. S., Nguyen, K. A., Stewart, R. A., & Giurco, D. (2020). Machine learning and data analytic techniques in digital water metering: A review. Water, 12(1).
- Nofal, S., Alfarrarjeh, A., & Abu Jabal, A. (2022). A use case of anomaly detection for identifying unusual water consumption in Jordan. Water Supply, 22(1).
- Qian, K., Jiang, J., Ding, Y., & Yang, S. (2020). Deep learning based anomaly detection in water distribution systems. In 2020 IEEE International Conference on Networking, Sensing and Control (ICNSC).
- Merta, J., & Fikejz, J. (2019). Utilization of machine learning to detect sudden water leakage for smart water meter. In 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA).

THANKS FOR WATCHING