Underground water pipes deteriorate under the influence of various physical, mechanical, environmental, and social factors. Reliable pipe failure prediction is essential for a proactive management strategy of the water supply network (WSN), which is challenging for the conventional physics-based model. This study employed data-driven machine learning (ML) models to predict pipe failures by leveraging the historical maintenance data and topographical datasets. The framework was developed for the Cleveland WSN dataset, which was found to achieve the best performance. The relative importance of major contributing factors on the water pipe failures was analyzed. Interestingly, the socio-economic factors of a community were found to affect the probability of pipe failures. This study indicates that data-driven analysis that integrates the Machine Learning (ML) techniques and proposed data fusion framework has the potential to support reliable decision-making in WSN management.

Figure 1. Overview of the Cleveland WSN network

Figure 2. Schema of aggregation of multiple sources of datasets

Figure 3. Illustration of assumptions used for data aggregation process of different datasets

Figure 4. Workflow for ML training and interpretation

Figure 5. Model summarization for pipe break prediction

Figure 6. The impacts of physical factors

Figure 7. Environmental factors impact

Figure 8. Impacts of operational factors

Figure 9. Impacts of community factors

Conclusion

- It provides a state-of-the-art data aggregation framework that integrates multi-source public datasets, leading to the largest real-world dataset (both in size and timeline) and associated largest number of input parameters for machine learning modeling.
- The LightGBM model achieved the highest performance with the second shortest training time.
- The SHAP interpretation results fit with previous studies, which demonstrated its ability to interpret the influence of studied factors. The results indicate the significant influence of pipe buried time, especially the interval time to the last break, experienced cold days, hot days, and pipe age.

Table 1: Performance evaluation of different ML models

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy in general</th>
<th>Speed of learning process</th>
<th>Handle categorical variables</th>
<th>Inherent model interpretability</th>
</tr>
</thead>
<tbody>
<tr>
<td>LightGBM</td>
<td>**</td>
<td>*</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>ANN</td>
<td>***</td>
<td>*</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>LR</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>kNN</td>
<td>*</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>SVC</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

*** denote the best performance
* denote the worst performance