XVIII World Water Congress
International Water Resources Association (IWRA)

Feature selection of acoustic signals for leak detection in water pipelines.

Ziyang Xu[a], Haixing Liu[a,*], Guangtao Fu[b], Yukai Zeng[a], Yunchen Li[a]
a School of Hydraulic Engineering, Dalian University of Technology, Dalian, Liaoning, 116024, China
b Centre for Water Systems, University of Exeter, Exeter, EX4 4QF, UK

Objectives

Water pipelines often have small cracks and leaks over time due to various degradation process. Acoustic emission technique is an outstanding method in leak detection for low cost and carrying out easily. However, the existing leak detection methods using acoustic signals primarily focused on identifying the crucial features, but take a limited consideration on the impact of feature interactions on leak detection. To address this gap, this study introduces a generalized feature selection framework called Maximal Discernibility and Minimal Redundancy and Improved Sequential Floating Forward Selection (MDMR_ISFFS). Five classifiers (i.e. DT, RF, XGBoost, SVM, and MLP) are used to examine the performance of the feature extracted from acoustic signals by using MDMR_ISFFS method through real water pipeline acoustic signals. Additionally, SHapley Additive exPlanations (SHAP) is utilized to analyze and elucidate the interaction mechanisms of the identified key features. The results demonstrate that four key features (i.e., Mean of frequency, Zero-crossing Rate, Peak frequency and Mean) are identified as the crucial features consistently regardless of the classifiers. All five classification models using MDMR_ISFFS can achieve high leak detection accuracies, ranging from approximately 94% to 98%. When compared to other feature selection methods (i.e., DFS_SFFS, KL distance and original feature set), the proposed MDMR_ISFFS can achieve the highest accuracy with a smaller number of features. Moreover, the study reveals significant interactions among the four key leakage features. In summary, this research provides valuable insights for selecting key leak features in actual pipeline leak detection.

Methods

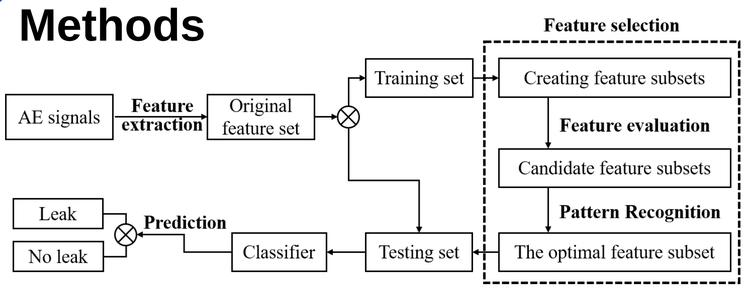


Fig. 1 Flowchart of the proposed methodology for leak detection. The AE signal-based leak detection is mainly followed by the three steps: (1) feature extraction; (2) feature selection; (3) prediction based on data-driven model.

Feature extraction is a fundamental and critical step to achieve data compression in the leak detection. In this study, we have chosen 17 time-domain features (referred to as T1-T17) and 7 frequency-domain features (designated as F1-F7) for analysis. Due to space limitations, specific feature extraction formulas are not provided here.

➢ Maximal Discernibility and Minimal Redundancy (MDMR)

- Discernibility of Feature Subset:

$$DFS = \frac{\sum_{i=1}^{n}\left(\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2\right)}{\frac{1}{N_+ - 1}\sum_{k=1}^{N_+}\left(\sum_{i=1}^{n}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2\right) + \frac{1}{N_- - 1}\sum_{k=1}^{N_-}\left(\sum_{i=1}^{n}\left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2\right)}$$

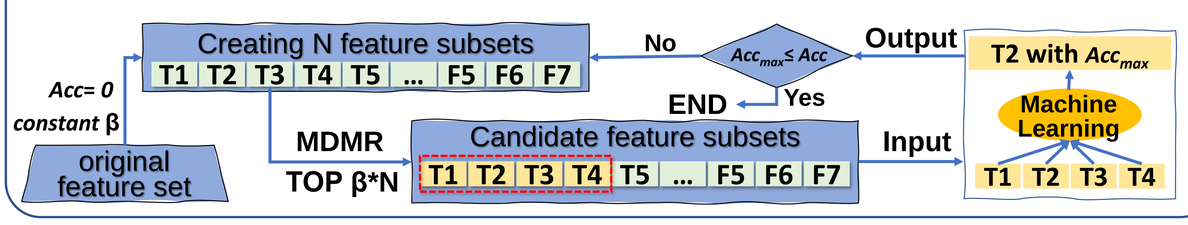
- Mutual Information:

$$RFS = \frac{1}{n^2}\sum_{x_i, x_j \in S} I(x_i; x_j)$$

$$MDMR(DFS, RFS) = \partial DFS - (1 - \partial) RFS$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ are the average of the ith feature of the whole feature subset, positive sample and negative datasets respectively; $x_{k,i}^{(+)}$, $x_{k,i}^{(-)}$ are the ith feature of the kth positive and negative cases respectively. The larger the MDMR value, the better the performance of the feature subset.

➢ Improved Sequential Floating Forward Selection (ISFFS)



Results

Table 1 The selected features and classifier performance on the test dataset. Using MDMR_ISFFS, all five models achieve high leak detection accuracies, approximately 94% to 98%. Notably, F1 (Mean of frequency), F5 (Peak frequency), T1 (Mean), and T14 (Zero-crossing rate) emerge as key features for leak detection, chosen by at least four classifiers.

Classifier	coefficients	Key Feature selected	Acc	Sen	Spe	F1 score
DT	∂=0.6; β=0.3;	F1; F5; T1; T14;	0.96	0.94	0.97	0.95
RF	∂=0.6; β=0.3;	F1; T14; T1; F5;	0.98	0.96	1.00	0.98
XGBoost	∂=0.3; β=0.5;	F1; T14; T1; F5; F6	0.98	0.94	1.00	0.97
SVM	∂=0.7; β=0.5;	F1; T14; T15; F5; T1; T11; F6; F2;	0.95	0.92	0.96	0.94
MLP	∂=0.6; β=0.7;	F1; T14; T13; F5; F2; T11;	0.94	0.94	0.94	0.93

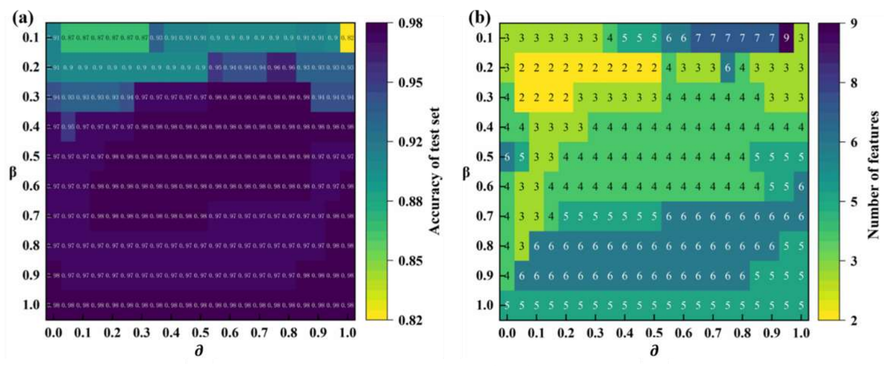


Fig. 2 Variation in leak detection Accuracy (Fig.2a) and the number of features selected (Fig.2b) under Different Combinations of Parameters ∂ and β

Accuracy of leak detection: Low β (≤0.3) leads to low model accuracy, indicating missing key features in candidate subsets. As β increases, model accuracy stabilizes on the test set, indicating successful interception of optimal subsets.
Number of features selected: After stabilization (β>0.3), higher ∂ leads to more optimal subsets, suggesting potential redundancy due to excessive weight on DFS.

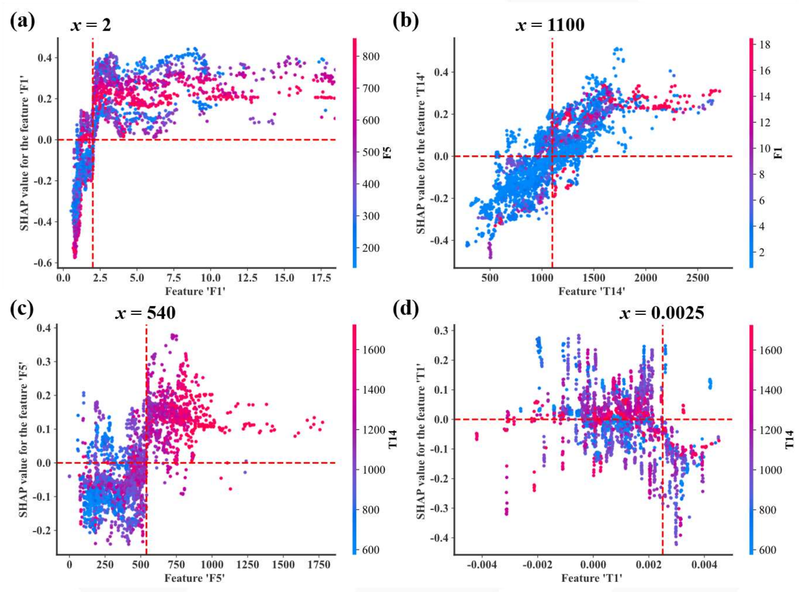
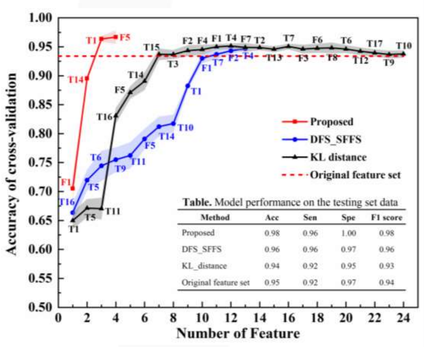


Fig. 3 Interaction between the key features impacts on the model output. These figures elucidate the predictive outcomes for training cases, with each data point corresponding to an individual case. X-axis represents feature values, while the Y-axis denotes respective SHAP values. Coloration indicates the values of interacting features, with red signifying high values and blue indicating low values.

Fig. 3 (a) shows that for F1 < 2, higher F5 reduces leakage probability. With increasing F1, higher F5 raises leakage probability. Fig. 3 (b) shows that for T14 < 1100, higher F1 decreases leakage probability. As T14 increases, F1 increases leakage probability. All the remaining graphs reach similar conclusions.

Fig. 4 Feature selection process and performance of different feature selection methods under RF classifier. The proposed RF-based feature selection method achieved 98% accuracy with 4 features, 96% with 13 features using DFS_SFFS, and 97% with 12 features using KL_distance. However, as feature count increased, performance stagnated or declined. Using the original feature subset directly yielded only 95% accuracy and a 94% F1 score on the test set. The paper also evaluates four other classifiers using different feature selection methods, all yielding consistent results, not detailed here.



Conclusions

In our paper, we introduce MDMR_ISFFS for feature selection and evaluate its performance in conjunction with five classifiers (DT, RF, XGBoost, SVM, and MLP) for real water distribution system leak detection. Here are the key conclusions:

- **High Accuracy**: All five classification models, when combined with our feature selection method, achieve impressive leak detection accuracies ranging from approximately 94% to 98%.
- **Key Features**: We identify four key features essential for leak detection. These are F1 (Mean of frequency), F5 (Peak frequency), T1 (Mean), and T14 (Zero-crossing rate). Each of these features is selected by at least four classifiers.
- **Feature Interaction Analysis**: Using the SHAP method, we analyze the interaction mechanism of these key features. High values of F1, T14, and F5 positively influence the predicted leakage probability (indicated by positive SHAP values), while a high value of T1 has a negative impact (negative SHAP value). Additionally, F5 and F1 have the strongest interactions with features F1 and T14, respectively, and T14 exhibits the strongest interactions with features F5 and T1.
- **Efficient Feature Selection**: Our proposed feature selection method converges quickly to achieve high accuracy with a smaller number of features compared to other methods.