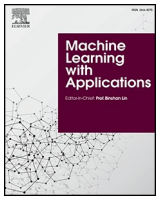
Machine Learning with Applications 14 (2023) 100501

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa



Semi-supervised anomaly detection methods for leakage identification in water distribution networks: A comparative study

Hoese Michel Tornyeviadzi *, Hadi Mohammed, Razak Seidu

Smart Water Lab, Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology, Ålesund, Norway

ARTICLE INFO

Keywords:
Anomaly detection
Leakage detection
Semi-supervised learning
Water distribution networks

ABSTRACT

This study presents a comprehensive evaluation of 10 state of the art semi-supervised anomaly detection (AD) methods for leakage identification in water distribution networks (WDNs). The performances of the semi-supervised AD methods is evaluated on LeakDB, a benchmark consisting of independent leakage scenarios that also account for the various sources of uncertainties arising in WDNs. Three performance metrics (F_β Measure, PR AUC Score, and Identification Lag Time) that collectively capture the different facets of leakage identification in WDNs is utilised to measure the efficacy of semi-supervised AD methods. Additionally, the TOPSIS MCDM tool supported with two weighting approaches is implemented to simultaneously consider all performance metrics in ranking the performance of semi-supervised AD methods. The results of this extensive comparative study shows that Local Outlier factor (LOF) is the overall best performing semi-supervised AD method on LeakDB. It is also evident that proximity based semi-supervised AD methods are superior to linear and probabilistic AD methods due to their ability to unearth leak events in the neighbourhood of normal operational data points. Finally, the impact of uncertainties on the performance of the semi-supervised AD models is discussed in addition to general recommendations on the usage of semi-supervised AD methods in leakage identification.

Introduction

Leakages are inevitable in water distribution networks (WDNs) mainly due to the deterioration of pipeline integrity. In most WDNs, pipelines are decades or centuries old making them exceptionally susceptible to structural integrity failure. Additionally, renewal rate of pipes is extremely low, approximately 1 % (EurEau, 2017). Ageing infrastructure coupled with low renewal rate has contributed to the persistence and recurring nature of leakages in WDNs. Leakages does not only result in gigantic volume of water loss, 126 billion cubic metres globally amounting to $39 billion per year (Liemberger & Wyatt, 2019), it also provides suitable avenue for contaminant intrusion under low pressure conditions which have severe public health implications (Besner et al., 2011). It is therefore imperative to actively look for leakages and promptly fix them to limit their adverse impact.

Proactive leakage management which entails seeking and finding leaks before they surface has been identified as the best management strategy for reducing leakages in WDNs (Jenks & Papa, 2022). Leakage detection in this context is accomplished in three (3) main steps namely, leakage identification which connotes the determination of the leak incident time, leak localization concerned with localizing and pinpointing the leak originating pipe and leak repair which deals with physically sealing the leak. The first step in leakage detection, leak identification which helps reduce the overall leak runtime significantly (AWWA, 2016) will be the focus of this study. Practically, reducing leak awareness time through leak identification is the most crucial since nothing can be done to resolve a leak if the utility is not aware of its occurrence. Leak localization and leak repair do not have significant impact on the overall leak runtime as compared to delayed leak identification. Their contribution to the overall leak runtime depends largely on the management principles adopted by the water utility (AWWA, 2016).

Over the last decade, several methods have been reported in literature for leakage identification with machine learning approaches dominating according to Wu and Liu (2017). Machine learning (ML) approaches to leak identification can be grouped into two categories: supervised and semi-supervised methods. Supervised ML methods have dominated the literature on leakage identification in recent years

* Corresponding author.
E-mail address: hoese.m.tornyeviadzi@ntnu.no (H.M. Tornyeviadzi).

https://doi.org/10.1016/j.mlwa.2023.100501
Received 1 September 2023; Received in revised form 22 September 2023; Accepted 26 September 2023
Available online 1 October 2023

(Hashim et al., 2020). Numerous supervised binary classification algorithms (e.g., Support Vector Machines, Naïve Bayes Classifier, and Multilayer Perceptron) have been reported in literature for leakage identification as highlighted by reviews from Wu and Liu (2017), and Hu et al. (2021). Supervised ML methods require accurately labelled data (leak and non-leak data) in WDNs for implementation. However, leaks in real-life WDNs cannot always be accurately labelled due to the fact that background leakages are inherent in Supervisory Control and Data Acquisition (SCADA) data, and the start time of a leak is usually not known explicitly. This hinders the practical utilization of supervised ML approaches in WDNs due to lack of labelled data, especially in newer WDNs that lack adequate historical data.

Additionally, supervised ML methods suffer from class imbalance. Class imbalance occurs when a classification dataset has skewed class proportions. In WDNs, the normal operation data representing the non-leak state far outweighs the leak event data owning to leakages being rare events. In order to overcome the class imbalance and lack of labelled data problem in leak identification, Villa-Pérez et al. (2021) posit the potential of semi-supervised anomaly detection (AD) methods in alleviating the shortcomings of supervised ML methods.

In semi-supervised learning, models are trained on only the majority class (normal operational data or "leak free" data) with the ultimate goal of detecting deviations from this majority class which represent the minority class (anomalies or leak events). The utilization of semi-supervised anomaly detection methods is gradually gaining momentum in literature for leakage identification in WDNs. Cody and Narasimhan (2020) presented a multivariate gaussian mixture model (GMM) model that utilizes linear prediction (LP) theory for multivariate SCADA data processing. Principal Component Analysis (PCA) coupled with Mahalanobis distance has also been widely utilised for leakage identification (Nam et al., 2019; Santos-Ruiz et al., 2018). Other semi-supervised ML approaches based on One Class Support Vector Machines (OCSVM) (Ayadi et al., 2019), Local Outlier Factor (LOF) (Muniz Do Nascimento & Gomes-Jr, 2022), and k Nearest Neighbours (kNN) (Vercruyssen et al., 2018) have also been reported in literature. The above studies have recognized the need to transition from supervised ML approaches to semi-supervised anomaly detection approaches for leakage identification in WDNs.

In terms of comparative studies, Vercruyssen et al. (2018) presented a comparison between five (5) semi supervised AD ML methods and expert identification of leakages. Recently, Muniz Do Nascimento and Gomes-Jr (2022) compared three (3) semi-supervised AD methods with expert annotation of leak events. Both studies emphasised semi-supervised AD methods present an automated compelling alternative to expert leak annotation in leak identification. Despite the progress presented by the aforementioned studies, some noteworthy limitations still exist. There is no general guidelines on the utilization of semi-supervised AD methods for leak identification and majority of these comparative studies evaluated only a small number of classical semi-supervised AD methods on limited datasets that do not account for a wide range of uncertainties in WDNs. In recent years, semi-supervised AD methods such as Fast Angle Based Outlier Detector (FastABOD) (Kriegel et al., 2008), and Copula-based outlier Detector (COPOD) (Li et al., 2020) have been making waves in the AD field. It has therefore become necessary to evaluate both the classical and these promising semi-supervised AD methods on extensive datasets that account for majority of uncertainties in WDNs to comprehensively evaluate their performance in leakage identification.

Additionally, the performance of semi-supervised AD methods in leakage identification were evaluated by previous studies using either Accuracy, F1score or Area Under the Receiver Operating Characteristic Curve (ROC AUC) Score. Accuracy and ROC AUC Score assume balanced classes which is not ideal for the class imbalance problem posed in leakage identification. Even though F1Score accounts for class imbalance, it assumes both precision and recall have equal importance or cost. In leakage identification systems, false positives have huge cost as compared to false negatives. False positives have the potential to erode the confidence of water engineers in leakage identification systems due to false alarms and also increase carbon footprint of leakages (leak repair crews drive to field in search of non-existent leaks). As such, the authors argue that precision should be given prominence over recall in the evaluation of semi-supervised AD methods.

Owing to these shortcomings, this study presents a comprehensive evaluation of ten (10) state of the art semi-supervised anomaly detection methods, including both classical (PCA, OCSVM and KNN) and promising AD methods (FastABOD and COPOD), in leakage identification. The AD methods were evaluated on LeakDB (Vrachimis & Kyriakou, 2018), a leakage diagnosis benchmark consisting of independent leakage scenarios that simultaneously account for three (3) different types of uncertainties in WDNs. Unlike previous studies that utilised single performance evaluation metrics that do not weigh precision appropriately or account for class imbalance, this study utilises three (3) performance metrics namely F_β measure, Area Under the Precision Recall Curve (PR AUC) Score and Identification Time Lag simultaneously to evaluate the performance of the AD methods. F_β measure permits assigning more weight to precision over recall and (PR AUC) is superior to ROC AUC on imbalanced datasets (Saito & Rehmsmeier, 2015). A single performance metric is incapable of accurately evaluating the different facets of leak identification in WDNs.

In order to fairly compare the performance of the semi-supervised AD methods, statistical comparison tests are conducted to analyse the differences in their performance and hierarchical clustering was used to ascertain groups of different semi-supervised AD methods with similar performance. The overall performance ranking of the semi-supervised AD methods is achieved through Technique for Order of Preference by Similarity (TOPSIS), a multi-criteria decision method (MCDM) tool that simultaneously considers all three performance metrics. TOPSIS ensures that each performance metric contributes its quota towards the ultimate ranking of the semi-supervised AD methods. Furthermore, unlike previous comparative studies that do not account for the different sources of uncertainties in WDNs, this study evaluates the impact of increasing uncertainty in the form of uncertainty in pipe length, pipe diameter and pipe roughness calibration on the performance of trained semi-supervised AD models.

Finally, general guidelines and recommendations on the utilization of semi-supervised AD methods in leakage identification is presented which is seldom considered by previous studies on leakage identification in WDNs. To the best of our knowledge, this comparative study is one of the first attempts to present a broad comprehensive comparative study in terms of the number of semi-supervised AD methods and provide some guidelines on the utilization of these methods in leakage identification systems in WDNs.

The study is organised as follows; Section 1 presents the introduction, Section 2 details the methods and materials. Specifically, it presents details on LeakDB, the semi-supervised AD methods, model performance metrics, statistical tests for model performance comparison and the TOPSIS MDCM tool. In Section 3, the results and discussions are presented. Finally, Section 4 presents the summary of the results, recommendations, limitations, and future work.

Materials and methods

This section presents the overview of the framework utilised to evaluate the performance of semi-supervised anomaly detection algorithms for leakage identification in WDNs. It commences with a brief description of the benchmark dataset used. This is followed by the elucidation of various state of the art semi-supervised anomaly detection methods for leakage identification and model performance evaluation metrics. Finally, a multicriteria decision making tool, TOPSIS, is presented to rank the performance of the anomaly detection methods in leakage identification. The overview of the entire methodology is presented in Fig. 1.

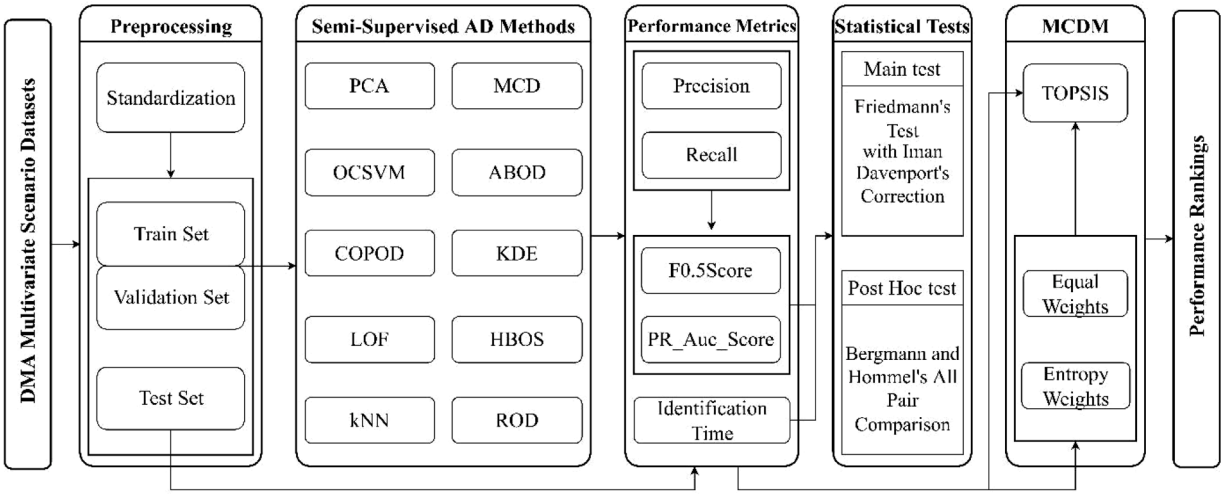


Fig. 1. Methodological framework.

Benchmark dataset description

Leakage Diagnosis Benchmark (LeakDB) (Vrachimis & Kyriakou, 2018) is a benchmark dataset for leakage diagnosis in WDNs. The dataset consists of both leakage and non-leakage scenarios with varying complexities on the Extended Hanoi WDN. The leakage scenarios include incipient leakages, abrupt leakages, and multiple leakages. When it comes to complexities in the benchmark dataset, the following are considered; uncertainty in pipe length, pipe diameter and pipe roughness calibration with maximum uncertainty level set to 25 %. In this study, only scenarios that meet the following criteria will be considered.

a) class imbalance ratio (IR) ≥ 11, and
b) uncertainty level ≤ 10 %.

The above-mentioned criteria is considered due to the fact that leakages are rare events, especially abrupt leakages, with relatively short leak runtimes (AWWA, 2016) resulting in huge class imbalance in real-life WDN SCADA data. The minimum class imbalance ratio is specified in order to consider only leakages of runtime less or equal to a month due to proactive leakage management in most water utilities. In terms of uncertainty levels, a maximum level of 10 % is enforced to prevent data or distribution shift in the scenarios to be evaluated. Uncertainties levels greater than 10 % completely changes the underlying distribution of the data. As such, these scenarios are discarded to ensure uniformity in the curated scenarios. In total, 66 independent scenarios from LeakDB that have met the above criteria were utilized for this study. See Table 4 in Appendix 1A for detailed information on these independent scenarios.

Multivariate data curation

A virtual district metering area (DMA) (Mamo et al., 2014) based multivariate data is curated for the identification of leaks within the Extended Hanoi WDN. In order to curate this multivariate data, the authors adopt the methodology reported in Steffelbauer and Fuchs-Hanusch (2016) to select the optimal pressure measurement points (nodes). Three (3) measurement points (Nodes 3, 14 and 24) were selected in addition to flow in Link 1. The selection of flow in link 1 is based on the design principles of WDNs. For each scenario, the dataset spans January 2017 to December 2017 with a frequency of 30 min. It is important to highlight that in the absence of physical DMAs in WDNs, virtual DMAs (Mamo et al., 2014) consisting of flow and pressure measurement points can be created. The utilization of multivariate data ensures robustness against single sensor data corruption. Additionally, it also helps with leak localization since each sector or DMA has a unique multivariate data. Therefore, leaks are confined to each (virtual) DMA or sector under ideal conditions. The smaller the sector or (virtual) DMA, the better the localization effort since the search area of the leak is significantly smaller as compared to gigantic physical DMAs in WDNs.

Data pre-processing

First, the multivariate SCADA data is standardized via z-scores. A z-score denotes standardization (Roshan et al., 2019) is adopted. The z-score denotes how many standard deviations a data point is above or below the mean. In the z-score standardization, the measurements representing leak events are amplified due to the fact that z-score is very sensitive to outliers which bodes well for leakage identification. The z-score is computed using Eq. (1), where x_i represents the raw sensor measurements, $\mu =$ mean, $\sigma =$ standard deviation and $z_i =$ standardized score.

$$z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

Semi-Supervised anomaly detection models

Semi-supervised anomaly detection methods belong to a family of machine learning algorithms that detect anomalies or novelties in unseen data by learning from only the normal/majority class of data (Chandola et al., 2009). Three (3) broad categories of semi supervised anomaly detection methods will be considered in this study. The broad categories include linear methods, probabilistic methods, and proximity-based methods. The subsequent subsections will expatiate on these broad categories and briefly highlight their associated anomaly detection methods. Detailed presentation of these methods can be found in their respective references.

Linear models

Linear models represent anomaly detection methods that encompass a linear combination of features in multivariate data to detect anomalies. From a semi-supervised perspective, the linear methods considered include Principal Component Analysis (PCA), Minimum Covariance Determinant (MCD), and One-Class Support Vector Machines (OCSVM).

PCA: Principal Component Analysis (PCA) is a technique commonly used for dimension reduction in multivariate data. It projects data into a lower dimensional space and also identifies the principal components that explain variability in the data (Shyu et al., 2003). It can be used for semi-supervised anomaly detection. In this case, a predictive model is constructed based on the major and minor principal components of the normal data. The anomaly score is calculated as the distance of an

anomaly from the normal data in the principal component space (Shyu et al., 2003).

MCD: Minimum Covariance Determinant (MCD) is a robust estimator of multivariate location (Rousseeuw & Driessen, 1999). When MCD is utilized for anomaly detection, it seeks to estimate the covariance matrix of the normal multivariate data such that the determinant is minimal. In the MCD implementation, the estimated covariance matrix is utilized in conjunction with the Mahalanobis Distance (MD) (Mahalanobis, 1936). The MD metric which represents the distance between a point and a distribution characterized by the covariance matrix serves an anomaly score.

OCSVM: One Class Support Vector Machine (OCSVM) (Schölkopf et al., 2001) utilizes either linear or non-linear kernel functions to map the input data into a high-dimensional feature space and determine the hyperplanes (decision boundaries) that best separate the points from the origin. Each new data point is classified based on the normalised distance from the decision boundary. Thus, test samples that are farther away from the delimited decision boundary are considered to be anomalous. The kernel functions typically utilised in OCSVM include linear, polynomial, radial basis, and sigmoid.

Probabilistic models

Generally, probabilistic-based models detect anomalies based on the probability score of a sample being anomalous. The semi-supervised probabilistic anomaly detection models considered include Angle-Based Outlier Detection (ABOD), Copula-Based Outlier Detection (COPOD) and Kernel Density Estimator (KDE).

ABOD: Angle-Based Outlier Detector (ABOD) (Kriegel et al., 2008) is an AD method for identifying anomalies in high dimensional data. It is headquartered on the variance of the angle between a new datapoint and all other pairs of observed datapoints. If the variance of the angle is low, the candidate point is considered an outlier otherwise it is a conforming datapoint. One major limitation of ABOD is its computational complexity arising from computing the variance of the angles between every pair of datapoint. In order to alleviate this problem, the authors of ABOD is proposed FastABOD . In FastABOD, only the *k*-nearest neighbors are considered.

COPOD: Copula-based outlier Detector (COPOD) is an AD detector inspired by copulas for modelling multivariate data distribution (Li et al., 2020). It is based on the non-parametric fitting of empirical cumulative distribution functions (ECDFs) called Empirical Copula to the normal multivariate data. First, it constructs the empirical copula, and then uses it to predict tail probabilities of each given new data point to determine whether its anomalous or not. COPOD is deterministic without hyperparameters and highly interpretable via quantifying the abnormality contribution of each dimension through dimensional outlier graph.

KDE: Kernel Density Estimator (KDE) is an AD method based on kernel density functions (Latecki et al., 2007). It adopts a modification of a nonparametric density estimate via a variable kernel to yield robust local density estimation on the normal data points. This local density estimate utilizes the reachability distance to enhance its robustness against large density estimates when data points are very close to their neighbours. Outliers are then identified by comparing the local density of each data point to the average local density of its neighbours. The balloon-type (Terrell & Scott, 1992) variable kernel is used in the density estimation.

Proximity based models

Proximity based models are headquartered on close proximity between data points. Data points outside this proximity range are considered anomalous. Thus, an anomaly depends on the degree of how isolated a data point is in relation to the surrounding neighbourhood. The following semi supervised proximity-based anomaly detection methods are considered in this study; Local Outlier Factor (LOF), Histogram-Based Outlier Score (HBOS), k Nearest Neighbours (KNN) and Rotation-based Outlier Detection (ROD).

LOF: Local Outlier Factor (LOF) is a distance-based anomaly detection method. It is headquartered on the degree to which a sample is isolated with respect to the surrounding neighbourhood (Breunig et al., 2000). First, the k-nearest neighbourhood of a data point is computed. Utilizing this k-nearest neighbourhood, the local density is then estimated via the local reachability density (LRD). The final LOF score is computed by comparing the local reachability density of a data point with the local reachability densities of its k nearest neighbours. Conforming data points have LOF scores close to 1.0 whiles anomalous data points have higher LOF scores.

HBOS: Histogram-Based Outlier Score (HBOS) is an AD method that utilises histograms to detect anomalies in data (Goldstein & Dengel, 2012). This method commences with the generation of univariate histogram per feature column of the multivariate data. For each feature, the frequency of samples falling into each bin is used to estimate its density (height of the bins). The histograms are then normalized such that the maximum height is 1.0. The HBOS score is computed as a product of the inverse of the estimated densities assuming the features are independent. HBOS is computationally less expensive, thus linear time complexity.

KNN: k-nearest-neighbour (kNN) (Ramaswamy et al., 2000) is a distance-based AD method. This AD method is accomplished in two stages. In stage one, the k nearest neighbour of each data point is computed. Based on these k nearest neighbours, the anomaly score is calculated in stage two. The anomaly score of a data point depends on the average distance to all the k nearest neighbours. Euclidean distance is the most widely used distance metric. Finally, the data points are then ranked based on their distance to their respective k nearest neighbours. The topmost ranked data points representing points with higher distances are declared as anomalies.

ROD: Rotation-based Outlier Detection (ROD) is a parameter-free AD method that requires no distribution assumptions on the multivariate data (Almardeny et al., 2020). This method is based on decomposing the feature space of the multivariate data into different combinations of subspaces. In these subspaces, 3D-vectors, representing the data points per 3D-subspace, are rotated about the geometric median two times counter-clockwise using Rodrigues rotation formula. The anomaly scores for each data point (3D representation) is computed as function of the median absolute value (MAD) of the normal dataset.

Another class of AD methods known as ensembles methods, which represent a combination of two or more base AD models, can also be identified in literature. Even though these ensemble AD methods usually outperform single base AD models, they are sometimes plagued with gigantic model size, and computational cost. Additionally, it is not 100 % guaranteed that complex ensemble AD models will always outperform base AD models (Galar et al., 2011). The maximum number of base models to include in an ensemble is still a debate in literature (Krawczyk, 2016). To ensure fair comparison and keep the model size and computation cost down, only single, or base AD methods are evaluated in this study. It is important to alert the reader that the authors are not attempting to discredit ensemble methods but rather giving elucidation on why they are not considered in the performance evaluation of AD methods in leakage identification.

Model training, validation and hyperparameter tuning

This section briefly explains the model training regime adopted, the method used for hyperparameter tuning, the cross-validation process and how these models were implemented.

Model training regime

Semi-supervised training is adopted to train the AD methods. Thus, only scenario(s) depicting leak free incident is utilised in the training of the models. Specifically, Scenario 210 is utilised. This scenario comprise of no leakages (incipient, abrupt, or multiple) and moderate uncertainty

levels (5 %) in the complexities considered. These moderate uncertainty levels were used in the training set in order to evaluate model robustness to changes in the WDN and the need for model retraining if uncertainty levels increase significantly. In the event that the training data is slightly corrupted with leakages, especially background leakages, the trained semi-supervised AD method cannot identify the leakages inherent in the training data. However, all new emergent leaks in the SCADA data will be identified effortlessly.

Model implementation and hyperparameter tuning

The AD models were implemented in Python. Specifically, PyOD (Zhao et al., 2019) a comprehensive and scalable Python toolkit for Outlier Detection in multivariate data is utilized. In addition, NumPy, scikit-learn, Pandas and Matplotlib libraries were used. A contamination level of 1e-6 was specified in the training of the models. In PyOD, the contamination level measured on the scale (0,0.5) represents the proportion of anomalies in the training data set. Since there are no anomalies in the training set, the contamination level is set to be close to zero as possible. The hyperparameters of each of the 10 AD methods were tuned using a grid search. Table 5 in Appendix 1B presents the specific hyperparameters of each AD method utilised in this study.

Cross-Validation process

A five (5) stratified KFold cross validation is implemented to ascertain the performance of the trained models. Since semi-supervised training is adopted, each validation fold is padded with additional data meeting the criteria stipulated in Section 2.1. This ensures that the class imbalance ratio is maintained in the validation set similar to that of the test set which the models will be evaluated on. These validation sets were also used in tuning model hyperparameters.

Model performance metrics

The performance of the semi-supervised AD methods in identifying leakages in WDNs is evaluated using the confusion matrix and its associated derivative metrics (Tharwat, 2020). The specific metrics considered include $F_{0.5}$ Score, and PR AUC Score due to the class imbalance nature of leakage identification in WDNs. When dealing with class imbalance problems, the minority class is often the most important and performance metrics should be chosen in a way to overcome the bias posed by the dominant class. $F_{0.5}$ Score, and Precision-Recall (PR) AUC Score focus solely on the minority class thereby curtailing this bias. In addition to these metrics, the identification time lag of each leakage is also evaluated.

$F_{0.5}$ score

This metric is a member of the F_β family of metrics which represent the harmonic mean of precision and recall. Here, $\beta = 0.5$ indicating more weight is given to precision over recall (Johnson & Khoshgoftaar, 2019). In WDNs, false positives have severe consequences such as erasing the confidence of SCADA operators in leakage identification systems and increasing carbon footprint (leak repair crews drive to field in search of non-existent leaks). The $F_{0.5}$ Score is given as

$$F_{0.5}\ \text{Score} = 1.25 \times \frac{\text{Precision} \times \text{Recall}}{0.25 \times \text{Precision} + \text{Recall}} \tag{2.1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.3}$$

Where True Positive (TP) represents correctly predicted leakages. False Positive (FP) represents incorrectly predicted leakages. False Negative (FN) represents incorrectly predicted no leakages.

PR auc score

The PR AUC score is an alternative class imbalance performance evaluation metric computed as area under the precision recall curve. The PR curve is a line plot between Precision and Recall for different probability thresholds that makes it possible to assess the performance of a classifier on the minority class (Haibo & Yunqian, 2013). The PR AUC Score ranges between 0 and 1 where 0 is the worst score and 1 represents the best score. The higher the AUC-PR score, the better the AD method in identifying leakages. PR AUC Score is more informative than ROC AUC on classifications problems with high class imbalance (Saito & Rehmsmeier, 2015).

Identification time lag

The identification time lag represents the time lapse between the occurrence of a leak event to when it is detected by the AD model. t_0 represents start of leak event and t_d represents time of identifying the leak event. The identification time lag is measures in hours.

$$\text{ITL} = t_d - t_0 \tag{3}$$

The identification time lag gives an indication of how long the leak has been running before it has been identified. To quantify the leak magnitude (or volume) and prioritize which leak to fix in the case of multiple leaks, the comparison of flow pattern distributions (CFPD) method (Thienen & Vertommen, 2016) can be utilised.

Statistical comparison of AD methods

In order to compare the performance of the AD methods and show whether there exist significant differences amongst them, appropriate statistical tests must be conducted. The Friedmans test is a non-parametric multiple comparison test analogous to repeated-measures ANOVA that aims to detect significant differences in the performance of at least two algorithms on multiple datasets (Derrac et al., 2011). This test seeks to ascertain if there is significant difference in the performance (e.g., $F_{0.5}$ Score, PR AUC score, etc.) of the AD methods on all the scenarios considered in LeakDB. The Iman Davenport's correction (Iman & Davenport, 1980) to the Friedman's statistic will be implemented in this study. This correction ensures that the test statistic is not undesirably conservative (Demšar, 2006). The hypothesis for the Friedman multiple comparison test is formulated as follows:

Ho: All anomaly detection methods have equivalent performance (equal rank).

Ha: At least one anomaly detection method differ in performance (unequal rank).

A significance level ($\alpha = 0.5$) is specified in the decision-making process to either reject or fail to reject the null hypothesis of equivalent performance of anomaly detection methods. See Appendix 1C for details on the Friedmans test and Iman Davenports correction.

Rejection of the null hypothesis indicates there is difference in the performance of the AD methods across scenarios considered in LeakDB. A slight draw back of the Friedman test is that it is unable to identify precisely which AD method(s) have dissimilar performance. Most often, post-hoc tests that conduct $N \times N$ comparison are utilized to compare the AD methods against each other in order to ascertain which AD method(s) differ in performance. In this study the Bergmann–Hommel's (Bergmann & Hommel, 1988) $N \times N$ post-hoc test is implemented. All statistical tests are implemented using the scmamp (Calvo & Santafé Rodrigo, 2016) library in R.

TOPSIS for combined performance ranking

The statistical methods presented in the preceding section can only ascertain the difference in performance of the AD methods with regards to a single performance metric at a time. In this study, there (3)

performance metrics ($F_{0.5}$ Score, PR AUC score and Identification Time Lag) were considered. Therefore, there is the need to evaluate and rank the performance of the AD methods considering all three (3) performance metrics simultaneously. This represents a multi criteria decision making (MCDM) problem. MCDM is a powerful tool for evaluating and ranking of many interconnected and competing criteria (performance metrics) that must be considered simultaneously (Benítez et al., 2020).

TOPSIS (Hwang & Yoon, 1981) is one of the few MCDM methods that finds a seamless blend of the unique strengths of all possible criteria. Thus, all criteria contribute their quota towards the ideal solution where a weakness in one criterion is compensated for by the strength of another criterion. However, TOPSIS is not without limitations, it requires explicit criteria weights in order to rank the alternatives (AD methods) which are not accounted for in the original methodology. Two priority weighting approaches, equal weights, and entropy weights are implemented in study to overcome this limitation. Equal weights assumes all criteria have equal importance while entropy weights gives higher importance to criteria that exhibit high variability (Boafo-Mensah et al., 2021). Ultimately, integrating these weighting approaches into TOPSIS presents a complete evaluation methodology that ensures AD methods are ranked objectively devoid of selection bias. Fig. 2 presents overview of the TOPSIS framework. Refer to Appendix 1D for mathematical formulations of the TOPSIS MCDM methodology.

The TOPSIS procedure commences with the formulation of the decision matrix, where the rows represents the alternatives (AD methods) and the columns represent the criteria, performance metrics ($F_{0.5}$ score, PR AUC Score and Identification Time). For each semi-supervised AD method, the median performance metric given all scenarios is utilized. The median is robust against outliers (AD method performing exceptionally well on a single scenario and awfully bad on majority of the scenarios). The decision matrix is then normalized to eliminate inconsistencies with different measurement units (scales) thereby transforming these units (scales) into a common measure for easy comparison. The normalized decision matrix is weighted via appropriate weighting scheme(s). Afterwards, the weighted decision matrix is evaluated, and the alternatives (AD methods) are ranked based on proximity to the ideal solution. $F_{0.5}$ Score and PR AUC Score are maximized indicating increase in these metric values contribute positively towards the ideal solution. On the other hand, Identification time is minimized indicating decreased values of this metric has positive influence on the ideal solution (ultimate ranking of semi-supervised AD methods in leakage identification).

Results and discussions

This section presents the results and discussions on the performance of the AD methods in leakage identification. It commences with a detailed presentation on the performance metrics and statistical tests conducted on these metrics. Then, results pertaining to the overall TOPSIS ranking framework utilised to select the best performing AD method is presented. Finally, the impact and implications of complexities in the datasets on trained AD models is highlighted.

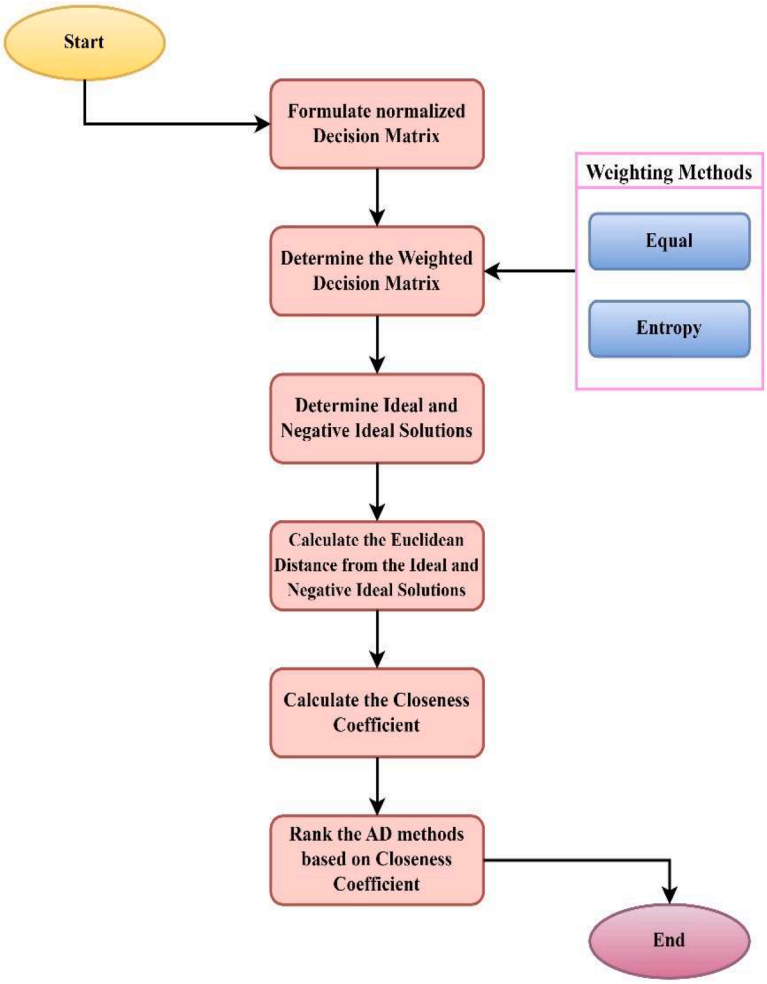


Fig. 2. Overall TOPSIS performance ranking framework.

Descriptive statistics on performance evaluation metrics

Fig. 3 represents box plots on metrics used to evaluate the performance of semi-supervised AD methods in leakage identification in WDNs. In this study, the metrics are grouped into two categories namely "ideal" and "common" metrics. The ideal metrics, $F_{0.5}$ Score, PR AUC Score and Identification Time Lag, represent the performance metrics adopted in his study whiles the common metrics; F_1 Score, Accuracy and ROC AUC Score, represent performance metrics utilised by prior studies. From Fig. 3, it is obvious that the ideal performance metrics ($F_{0.5}$ Score and PR AUC Score) which account for class imbalance do not overestimate the skill of the semi-supervised AD methods in leakage identification. On the other hand, Accuracy and ROC AUC Score which do not account for class imbalance significantly exaggerates the performance of the semi-supervised AD methods in terms of median performance across scenarios in LeakDB. Even though F_1 Score considers class imbalance, it slightly overemphasizes the skill of the semi-supervised AD methods due to assigning equal importance to precision and recall. Similar results have also been obtained in literature by other comparative studies (Bekkar et al., 2013; Saito & Rehmsmeier, 2015) on performance metrics for highly imbalanced problems further corroborating the findings of this study.

The rest of this study will focus exclusively on the ideal performance metrics (row 1 of Fig. 3) for the evaluation of semi-supervised AD methods. From the boxplots of $F_{0.5}$ Score and PR AUC score, it is clearly evident that the AD methods do not perform consistently across all the curated scenarios in LeakDB (existence of outliers). These outliers represent instances where the AD method(s) perform exceptionally well on particular scenario(s) whiles the average/median performance across all scenarios is significantly lower. PCA, OCSVM and KDE are the major culprits, whiles the other AD methods seem to exhibit appreciable level of consistency across all scenarios. A deep dive into the results reveal that AD methods such as PCA, OCSVM and KDE perform best when the uncertainty levels are minimal (≤ 5 %). Their performances deteriorate significantly once the uncertainty levels increase indicating these models do not seem to generalise well. ROD is the worst performing AD method since it failed to identify any of the leaks in LeakDB which is rather unexpected.

In terms of $F_{0.5}$ Score which places emphasis on precision over recall, LOF has the best performance (highest mean and median) amongst the semi-supervised AD methods. The mean and median also coincide, indicating the $F_{0.5}$ Scores are normally distributed over the scenarios in LeakDB. MCD and KNN also seem to perform well comparatively. In a nutshell, these AD methods resulted in less false positives. AD models that consistently produce less false positives are highly desirable in the proactive management of WDNs. This is because they eliminate needless and non-existent leak searches by the leak repair crew thereby significantly reducing cost and associated CO_2 footprint. When it comes to the PR AUC Score, MCD, ABOD, LOF and KNN seem to represent the best performing AD models. On average, these AD methods were able to identify the leak incidents in the various scenarios considered in LeakDB.

PCA, OCSVM, KDE and HBOS performed poorly with regards to the PR AUC Score performance assessment metric. Even though majority of these PR AUC Scores are very small, it is important to highlight the fact that these scores are still greater than the no skill PR AUC threshold. The no skill PR AUC threshold is given as the inverse of the class imbalance ratio (IR) (Saito & Rehmsmeier, 2015). Given all the scenarios curated in LeakDB, the class IR range is [11 472.5]. This implies the minimum PR AUC Score threshold is 2.1186×10^{-3} and the maximum is 0.0910. From the PR AUC scores boxplots in Fig. 3, it is obvious that majority of the values are higher than the maximum threshold, 0.0910, across all scenarios considered in LeakDB. The reader is reminded the boxplot presents an overview of the entire data, not scenario specifics tied to a particular PR AUC score threshold.

Identification Time lag is another crucial performance indicator in the evaluation of AD methods in leakage identification. It indirectly provides an indication of the volume of water lost to leakages once the leak magnitude is known or estimated. LOF and ABOD recorded the least identification time lag indicating prompt identification of leakage incidents. According to AWWA (2016), the anatomy of leak runtime include awareness (identification), localization, and repair. Most often leak localization and leak repair times are fixed depending on the practices adopted by the water utilities. The awareness time is what varies significantly. Therefore, prompt leakage identification plays a predominant role in reducing the overall leak runtime. The

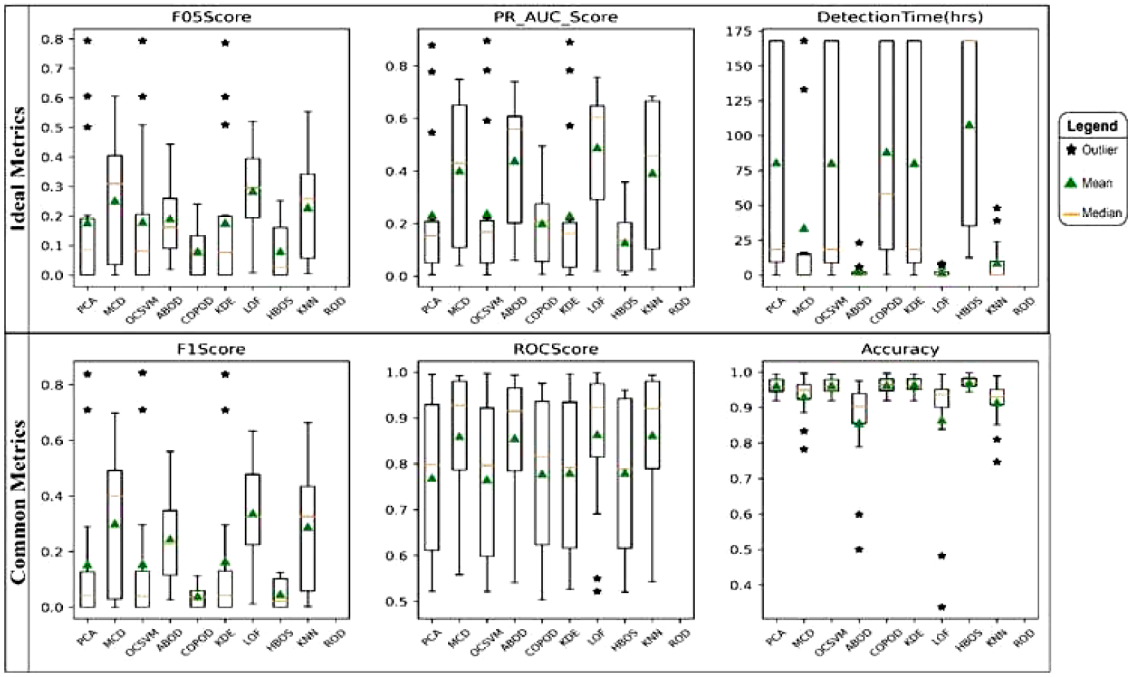


Fig. 3. Boxplot of class imbalance aware performance metrics vs traditional metrics.

identification time of the other AD methods, PCA, OCSVM, COPOD, KDE and HBOS, vary widely indicating less than stellar performance in prompt identification of leakages in WDNs.

Statistical comparison of performance evaluation metrics

Even though the boxplots presented in the previous section provide great visual clues into the performance of the AD models, it is unable to establish if there exist significant difference in the performance of the AD models. This subsection seeks to establish if there is any significant difference in the performance of the semi-supervised AD models through Friedmans test.

Table 1 presents the average ranks of the AD methods across all scenarios considered in LeakDB in addition to the results of the Friedmans Test with Iman Davenport's correction. In Table 1, ROD is omitted because it fails to identify any of the leaks (unavailability of performance metric values). LOF consistently has the best average ranking in all performance metrics, followed by MCD and KNN. This implies that LOF consistently outperforms the other AD methods on majority of the scenarios in LeakDB. There appear to be some degree of similarity in the average ranking in the different performance metrics. HBOS and COPOD attained the worst average ranking in all performance metrics. These AD methods exhibited poor performance (see Fig. 3) on the scenarios in LeakDB and therefore always ranked the least. In a nutshell, it is evident in Table 1 that there is significant difference in the performance of the AD methods as shown by the p-values < 0.05 for each performance metric.

Table 2 presents the results of the $N \times N$ Bergmann post-hoc test that seeks to ascertain the semi supervised AD method(s) responsible for the difference in performance. In these pairwise comparisons, there is significant differences in the $F_{0.5}$ Score between MCD vs. HBOS, COPOD vs. LOF, LOF vs. HBOS and HBOS vs. KNN. Similar significant differences are also exhibited in the performance metrics PR AUC score and Identification Time Lag. A critical look at Table 2 reveals that the difference in performance with regards to PR AUC Score seems to directly replicate in the Identification Time Lag column. A plausible explanation of this occurrence could be attributed to the relationship between PR AUC score and Identification Time Lag. PR AUC score only focuses on the minority class, thus the ability of the classifier (AD method) to accurately highlight leak events. A good PR AUC score (skill of the AD method to accurately highlight/tract leak events) will translate to prompt leak identification. As such, the difference in performance seems to be similar between these performance metrics.

TOPSIS based selection of overall best performing AD method

In this section, results on the ranking of the semi-supervised AD methods considering all three performance metrics ($F_{0.5}$ Score, PR AUC Score and Identification Time Lag) simultaneously is presented. The decision matrix in the TOPSIS ranking procedure is formulated using the median of each performance metric. The median is a robust measure of central tendency and therefore a good point estimate to represent the performance across all scenarios in LeakDB.

Table 3 presents the results of the ranking of the semi-supervised AD methods in leak identification in WDNs. LOF is ranked as the best

Table 1
Friedmans test results.

AD Method	Average Ranking		
	F0.5Score	PR AUC Score	Identification Time Lag
PCA	5.8000	6.5333	6.6000
MCD	3.2000	3.3333	3.6000
OCSVM	5.5000	5.9333	6.4000
ABOD	4.3333	3.0000	2.1333
COPOD	7.0667	5.8000	7.4667
KDE	5.8000	7.0000	6.1667
LOF	2.9333	2.3333	2.3333
HBOS	7.1667	7.6000	7.6333
KNN	3.2000	3.4667	2.6667
Test Statistic	7.927	14.829	31.735
p-value	2.068e-08	1.266e-14	$< 2.2e-16$

Table 2
Bergmann and Hommels post-hoc test results.

Serial	AD Methods	p values under performance metrics		
		F0.5Score	PR AUC Score	Identification Time Lag
1	PCA vs. MCD	0.149	**0.016**	**0.027**
2	PCA vs. OCSVM	1.000	1.000	1.000
3	PCA vs. ABOD	1.000	**0.007**	**0.000**
4	PCA vs. COPOD	1.000	1.000	1.000
5	PCA vs. KDE	1.000	1.000	1.000
6	PCA vs. LOF	0.091	**0.001**	**0.000**
7	PCA vs. HBOS	1.000	1.000	1.000
8	PCA vs. KNN	0.149	0.022	**0.001**
9	MCD vs. OCSVM	0.257	0.930	0.051
10	MCD vs. ABOD	1.000	1.000	1.000
11	MCD vs. COPOD	**0.002**	0.136	**0.001**
12	MCD vs. KDE	0.015	**0.004**	0.103
13	MCD vs. LOF	1.000	1.000	1.000
14	MCD vs. HBOS	**0.002**	**0.000**	**0.001**
15	MCD vs. KNN	1.000	1.000	1.000
16	OCSVM vs. ABOD	1.000	**0.044**	**0.000**
17	OCSVM vs. COPOD	1.000	1.000	1.000
18	OCSVM vs. KDE	1.000	1.000	1.000
19	OCSVM vs. LOF	0.164	**0.006**	**0.001**
20	OCSVM vs. HBOS	1.000	1.000	1.000
21	OCSVM vs. KNN	0.257	0.136	**0.002**
22	ABOD vs. COPOD	0.091	0.061	**0.000**
23	ABOD vs. KDE	1.000	**0.001**	**0.001**
24	ABOD vs. LOF	1.000	1.000	1.000
25	ABOD vs. HBOS	0.091	**0.000**	**0.000**
26	ABOD vs. KNN	1.000	1.000	1.000
27	COPOD vs. KDE	1.000	1.000	1.000
28	COPOD vs. LOF	**0.001**	**0.008**	**0.000**
29	COPOD vs. HBOS	1.000	1.000	1.000
30	COPOD vs. KNN	**0.002**	0.196	**0.000**
31	KDE vs. LOF	0.091	**0.000**	**0.002**
32	KDE vs. HBOS	1.000	1.000	1.000
33	KDE vs. KNN	0.149	**0.007**	**0.005**
34	LOF vs. HBOS	**0.001**	**0.000**	**0.000**
35	LOF vs. KNN	1.000	1.000	1.000
36	HBOS vs. KNN	**0.002**	**0.001**	**0.000**

Table 3
TOPSIS ranking of AD methods on scenarios in LeakDB.

Methods	**Median Values**			**Equal Weight TOPSIS (Rank)**	**Entropy Weight TOPSIS (Rank)**
	F0.5Score	**PR_AUCScore**	**Time Lag (hrs)**		
PCA	0.09	0.15	18.50	0.5880 (6)	0.8474 (5)
MCD	0.31	0.43	0.50	0.8737 (2)	0.9745 (2)
OSCVM	0.08	0.17	18.50	0.5892 (5)	0.8473 (6)
ABOD	0.16	0.56	0	0.7928 (4)	0.9411 (4)
COPOD	0.08	0.21	58.00	0.4952 (8)	0.6434 (8)
KDE	0.08	0.16	18.50	0.5846 (7)	0.8459 (7)
LOF	0.30	0.60	0	**0.9786 (1)**	**0.9943 (1)**
HBOS	0.03	0.14	168.00	0 (9)	0 (9)
KNN	0.26	0.46	0	0.8690 (3)	0.9705 (3)
ROD	–	–	–	–	–

performing AD method followed by MCD and KNN. COPOD and HBOS are the worst performing AD models on the scenarios in LeakDB. The TOPSIS rankings by the two priority weighting approaches are similar in Table 3. This similarity is further confirmed with a Kendall's rank correlation coefficient of 0.9444 and a p-value of 4.9603e-05. Thus, whichever weighting approach is adopted, the TOPSIS rankings remain consistent. However, if there is the need to place much emphasis on any of the performance metrics, thus assign much importance as compared to the other performance metrics, appropriate weighting scheme such as fuzzy AHP (Li & Li, 2009) could be adopted. Fuzzy AHP takes into consideration the opinion of expert engineers in the priority rating of performance metrics used to evaluate the AD methods. The min-mode-max fuzzy triangular number formulation of the priority ratings reported in our previous study (Tornyeviadzi et al., 2021) helps eliminate conflict in the priority rating by various experts with diverse expertise and backgrounds. In Table 3, ROD is not considered in the TOPSIS ranking due to the fact that it failed to identify any of the leaks in LeakDB.

In Fig. 4, the leak identification results on Scenario 004 is presented. The results in Fig. 4 partly confirm the TOPSIS rankings with LOF having perfect precision on this scenario. One prominent observation is how most AD methods misclassify data points in the summer as leak events. Generally, in the summer temperatures rises, consumption patterns change, and flow measurements become highly unpredictable due to rapid/instantaneous increases in consumption that persist for few minutes (Kara et al., 2016). This occurrence gradually results in marginal increase in flow accompanied with slight drop in pressure. Majority of the AD methods fail to make this distinction in the summer and therefore highlighted some of the data points as leak events. KDE is the worst culprit followed by OCSVM and PCA. A plausible remedy for this occurrence is to introduce artificial delay during the summer period. This artificial delay would require multiple successive time points before a leak alarm is triggered. In a nutshell, this remedy endeavours to lessen the negative impact of short lived rapid increases in consumption on leakage identification systems.

AD model type and leak identification performance

In order to identify natural clusters (similar model behaviour) in the performance of the semi-supervised AD methods with regards to the three-performance metrics simultaneously, hierarchical clustering (Murtagh & Contreras, 2012) is adopted in this study. Hierarchical clustering groups data into a multilevel cluster tree represented by a dendrogram. For each semi supervised AD model, its median performance metric is used as a proxy. Fig. 5 presents the natural clusters in the performance of the semi-supervised AD methods. Three clusters were found which coincidentally conforms to the ranking of the AD methods. Cluster 1 consists of the best four (4) methods namely LOF, MCD, KNN and ABOD. Cluster 2 consists of PCA, OCSVM and KDE. Finally, cluster 3 consists of COPOD and HBOS. Each cluster represents AD methods that have similar performances on LeakDB. It is imperative to understand what drives this natural clusters in the performance of the AD methods and its implications.

Cluster 1 represents the 1st hierarchical level of the multilevel cluster. A critical look at the AD methods (LOF, MCD, KNN and ABOD) attributed to cluster 1, reveals majority (except MCD) are proximity-based approaches that utilised the k nearest neighbours' method. Additionally, their anomaly score is calculated based on a distance metric. This implies proximity-based AD approaches seem to perform best for leak identification in WDNs compared to linear and probabilistic AD approaches. Leaks, especially abrupt leakages, result in sudden rise in flow accompanied by decrease in pressure. From a neighbourhood perspective, this will result in significant difference between normal operational data and leak event data. Proximity-based approaches are well suited to capture these changes in the neighbourhood of data points and highlight them as leakages.

In cluster 2, PCA and OCSVM are linear methods whiles KDE is a probabilistic method. However, KDE makes use of the linear median finding algorithm in its KD-tree implementation (Munaga & Jarugumalli, 2011). These group of AD methods represents the 2nd level in the hierarchy of the multilevel cluster. Their performance is not up to par with the proximity-based approaches due to the fact that they assume linearity in the multivariate data. Diurnal consumption patterns,

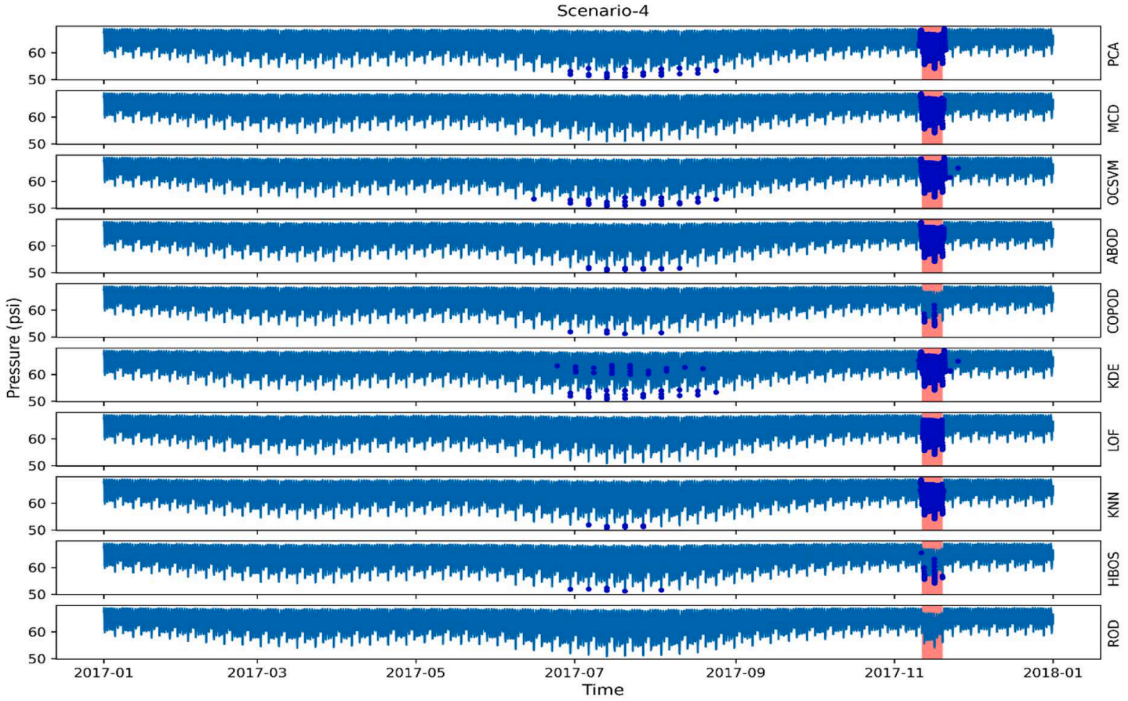


Fig. 4. AD methods performance on Scenario 004 in LeakDB.

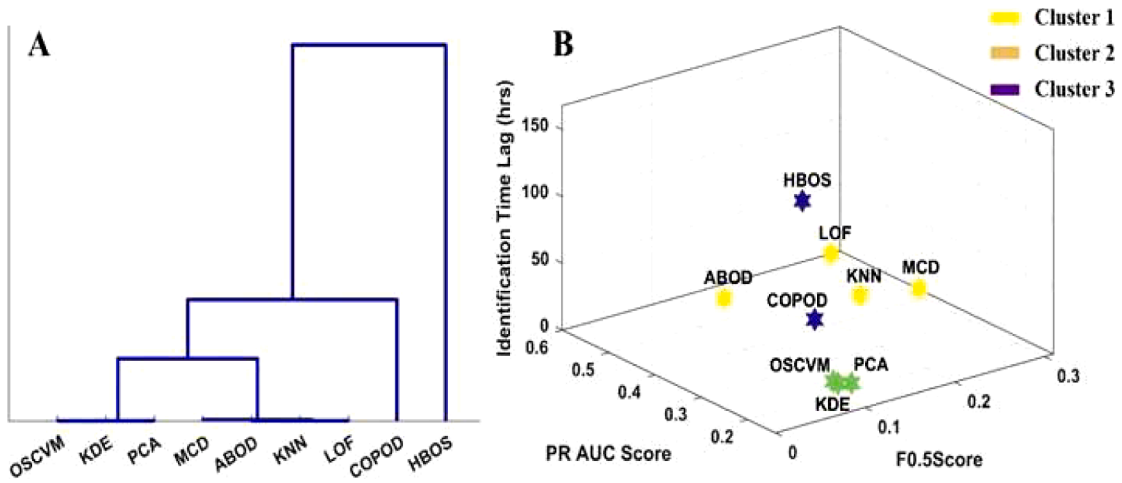


Fig. 5. Hierarchical multilevel clusters in AD method performance.

especially residential patterns, in WDNs are typically nonlinear. This makes 24 hr or weekly WDN SCADA data non-linear with their respective periodicities. However, over a long horizon, for example in a particular season, the trends could be linear which favours the implementation of linear AD methods. Since one year data is utilised in this study, these linear methods are able to identify some of the leak events.

Cluster 3 which represents the 3rd hierarchical level in the multilevel cluster comprises of COPOD and HBOS. These AD methods represent the worst performing models on LeakDB. This worst performance is attributed to the underlying assumptions associated with COPOD and HBOS. HBOS explicitly assumes each feature column (variable) of the multivariate SCADA data is independent (Goldstein & Dengel, 2012), whiles COPOD computes the joint probability distribution of the multivariate data using only the marginal probabilities. Thus, each dimension (feature column or variable) of the multivariate data is modelled separately (independently) and linked together to form the joint distribution through copulas (Li et al., 2020).

For any given DMA in a WDN, the relationship between flow and pressure is well documented. Additionally, in the event of a leak, pressure residuals dissipate in magnitude based on proximity from the leak originating pipe (Perelman et al., 2016) alluding to some level of collinearity between pressure sensors in a DMA or sector. These aforementioned realities make it difficult to decouple features from multivariate data and treat them separately or independently. As such, COPOD and HBOS performed poorly since their underlying model assumptions do not fit well with multivariate WDN SCADA data.

Impact of increasing uncertainties on AD methods

In this section the impact of increasing uncertainty in the form of changes in pipe roughness, pipe diameter and pipe length is evaluated. Uncertainty magnitudes greater than 10 % of the baseline on which the AD models were trained for all uncertainty categories are considered. Fig. 6 represents results on one of the numerous scenarios in LeakDB

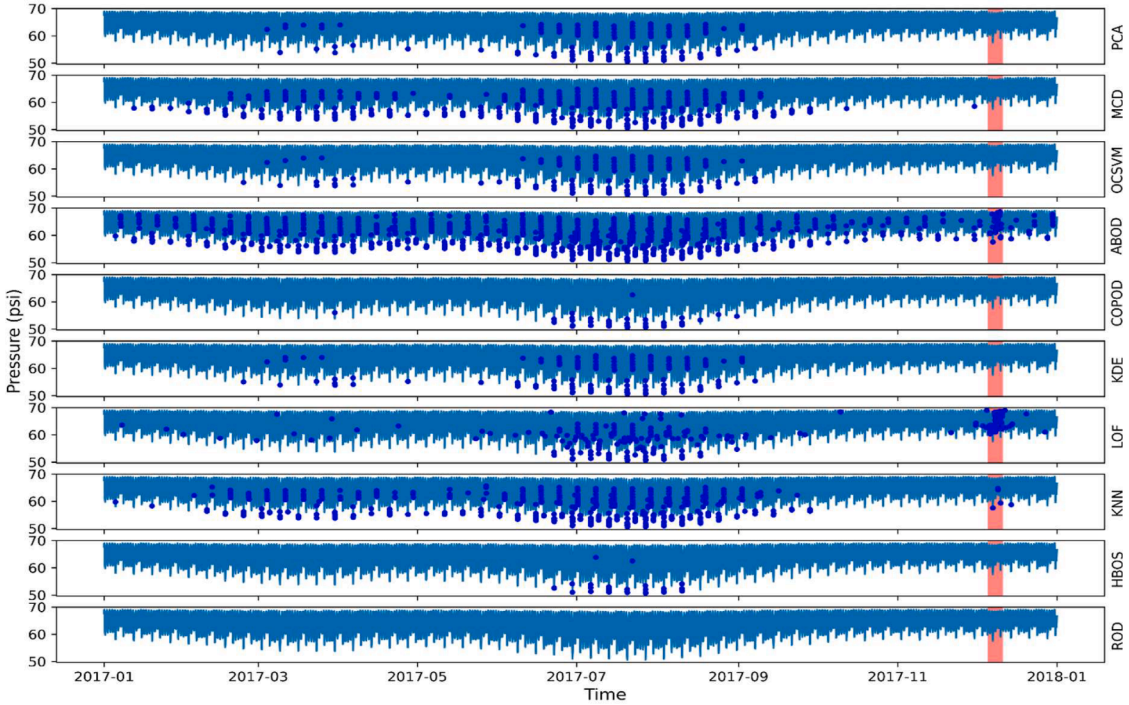


Fig. 6. Impact of increasing uncertainties on AD methods performance.

that has uncertainty magnitudes greater than 10 % across all uncertainty categories (pipe roughness, pipe diameter and pipe length). It is evident in this figure that the AD models perform poorly. ABOD has the worst performance. Majority of the AD methods (PCA, MCD, OCSVM, COPOD, KDE, HBOS, and ROD) actually fail to identify the leak event due to the increasing uncertainty. This increasing uncertainty has resulted in a major data drift, thus changes in the AD model input data in unforeseen ways, which is further confirmed with a maximum mean discrepancy (Gretton et al., 2012) test statistic of 0.9444 and a p-value of 4.9603e-05. The maximum mean discrepancy (MMD) is a non-parametric statistical test used to determine multivariate drift. As shown in Fig. 6, data drift or distribution shift ultimately translates to model performance degradation. Even though LOF and ABOD attempted to identify the leak events, numerous false positives have been recorded rendering their identification precision under increasing uncertainties useless.

The results in Fig. 6 also demonstrate that there is the need to monitor data drifts and distribution shifts in the implementation of data driven AD methods. This will ensure that swift actions such as model retraining is conducted in a timely fashion to mitigate the adverse impact of data drift on leakage identification. In real-life WDNs, a major short term contributing factor to increasing uncertainty (data drift or distribution shift) in SCADA data is partial rehabilitation of the pipe network after model deployment. Other factors such as ageing of pipes, and shift in consumption patterns also have the tendency to change the distribution of SCADA data. However, none of them contribute significantly in the short term as compared to rehabilitation.

Conclusions, recommendations and future work

This section presents the summary of findings, recommendations for the application of semi-supervised AD methods in leakage identification, limitations, and possible future directions of the study.

Summary of findings

The performances of ten (10) state of the art semi-supervised AD methods in leak identification in WDNs have been evaluated on LeakDB and ranked through the TOPSIS MCDM framework in this study. The following key points summarise the findings of this comparative study.

- Amongst the three (3) broad categories of semi-supervised AD methods (linear, probabilistic and proximity) considered in this study, proximity based approaches (LOF, KNN, ABOD) present the best performance on LeakDB. Specifically, LOF has the best performance in all three performance metrics with a median $F_{0.5}$ Score of 0.30, median PR AUC Score of 0.60 and median Identification time of 0 hrs.
- Linear AD methods such as PCA and OCSVM in addition to KDE represent the 2nd best performing group of AD methods. Their performance is not on par with proximity based approaches due to the linearity assumption in linear AD methods. COPOD and HBOS are the worst performing AD methods on LeakDB. Their model assumption of independent features are unrealistic for multivariate SCADA WDN data. ROD failed to identify any of the leak events in LeakDB.
- The TOPSIS MCDM tool has presented a comprehensive framework to rank the performance of the AD methods by considering all three performance metrics ($F_{0.5}$ Score, PR AUC Score and Identification Time Lag) simultaneously. Both equal and entropy weights incoporated into TOPSIS for criteria importance/priority weighting presented consistent rankings of AD methods performance. The TOPSIS framework provides a wholistic and all-inclusive performance evaluation for AD methods.
- Increasing uncertainties deteriorate semi-supervised AD model performance. In worst case scenario, they result in distribution shift (data drift). Uncertainties emanating from different sources (pipe length, pipe diameter and pipe roughness) have compounding effect when considered simultaneously. Greater care must be taken to limit the impact of uncertainties and always ensure model input data is close to training data utilised.

Recommendations

Based on the findings of this study the following general recommendations are made with regards to the utilization of semi-supervised AD methods for leak identification in WDNs.

- Due to the huge class imbalance in WDN SCADA data, appropriate performance metrics that give prominence to the minority class should be adopted. These performance metrics include, F_β Score with $\beta < 1$, and PR AUC Score. $\beta < 1$ is recommended due to the fact that there is a higher cost associated with false positives in leakage identification systems as compared to false negatives. Additionally, multiple performance metrics should be considered simultaneously to wholistically evaluate performance.
- Proximity based semi-supervised AD methods that utilise k nearest neighbours with distance based anomaly scores favour the identification of leakages, especially abrupt leakages. Intuitively, the emergence of a leak is usually coupled with increase in flow and drop in pressure under ideal conditions. This results in significant deviation in the neighbourhood of data points in the horizon of the leak. Proximity based approaches are best positioned to unearth these changes in the neighbourhood of the normal data points.
- Semi-supervised AD models should be retained when operational activities that has the tendency to change the model input data distribution or result in data drift are carried out in WDNs. Some of these operational activities include complete or partial rehabilitation of the network, drastic changes in consumer patterns, extension of network to new areas, introduction of new pressure regulating valves that alter the distribution of flow and pressure on the network significantly. These activities should trigger retraining of AD models to ensure consistency between expected model input and actual model input.
- Semi-supervised AD methods that decouple multivariate SCADA data in WDNs are not ideal for leakage identification. These AD methods disregard the inherent relationship between pressure measurements in the presence of a leak within a DMA by treating each feature column of the multivariate data independently. Methods based on this assumption, decoupling of features in multivariate data, do not represent the reality in WDNs and should not be favoured in leakage detection systems.

Limitations and future work

This study is not lacking limitations, the impacts of missing data and sensor drifts are not evaluated due to the absence of these features in the benchmark dataset. In real-life WDNs, installed sensors drift with time requiring recalibration and due to power outages sensors may also fail to transmit data to the central SCADA system on time. Sensor drifts and data pre-treatment such as interpolation of missing data further increases data uncertainty. Even though other forms of uncertainty have been evaluated in this study, uncertainties emanating from the treatment of missing data, and sensor drifts could be evaluated in future studies. Ensemble AD models that combine several base AD methods should also be investigated to assess the trade-off between model performance improvement and increased computational cost. Furthermore, contextual AD methods that utilise sequenced data such as sequence-to-sequence deep autoencoders should be investigated for leakage identification in conjunction with exploring other class imbalance performance assessment metrics.

CRediT authorship contribution statement

Hoese Michel Tornyeviadzi: Conceptualization, Methodology, Data curation, Formal analysis, Software, Visualization, Writing – review & editing. **Hadi Mohammed:** Conceptualization, Writing – review & editing. **Razak Seidu:** Supervision, Conceptualization, Funding acquisition, Resources, Project administration, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Benchmark dataset openly available.

Acknowledgments

This work was supported by the Smart Water Project (project no: 90392200) funded by NTNU Smart Water Lab and Ålesund Municipality.

Appendix 1A: Scenarios utilized in LeakDB

Table 4

Table 4
Leak scenario details.

Scenario id	Leak type
Scenario-004	Single abrupt leak
Scenario-034	Multiple leak
Scenario-038	Multiple leak
Scenario-114	Single incipient leak
Scenario-135	Single abrupt leak
Scenario-167	Single abrupt leak
Scenario-170	Single abrupt leak
Scenario-185	Multiple leak
Scenario-212	Single incipient leak
Scenario-246	Single abrupt leak
Scenario-282	Multiple leak
Scenario-331	Single abrupt leak
Scenario-367	Single abrupt leak
Scenario-403	Multiple leak
Scenario-423	Multiple leak
Scenario-426	Single abrupt leak
Scenario-435	Single incipient leak
Scenario-456	Single abrupt leak
Scenario-494	Multiple leak
Scenario-506	Multiple leak
Scenario-530	Single abrupt leak
Scenario-552	Single incipient leak
Scenario-557	Multiple leak
Scenario-652	Single abrupt leak
Scenario-684	Single abrupt leak
Scenario-717	Single abrupt leak
Scenario-725	Multiple leak
Scenario-739	Single incipient leak
Scenario-798	Single incipient leak
Scenario-827	Single incipient leak
Scenario-832	Single abrupt leak
Scenario-841	Multiple leak
Scenario-863	Single incipient leak
Scenario-905	Single incipient leak
Scenario-910	Multiple leak
Scenario-926	Single incipient leak
Scenario-944	Single abrupt leak
.	.
.	.
.	.
Scenario-210	No leak

Appendix 1B: Hyperparameters of AD methods

Table 5

Table 5
Hyperparameters of AD methods.

Methods	Hyperparameters
PCA	contamination = 1e-6; n_components = 2; svd_solver = 'full svd'
MCD	contamination = 1e-6;
OCSVM	contamination = 1e-6; kernel= 'poly'; gamma = 1/n_features;
ABOD	contamination = 1e-6; n_neighbors = 12
COPOD	contamination = 1e-6;
KDE	contamination = 1e-6; algorithm = 'kd_tree'; leaf_size = 48
LOF	contamination = 1e-6; n_neighbors = 12; algorithm= 'ball_tree'; leaf_size=48
HBOS	contamination = 1e-6; n_bins = 10
KNN	contamination = 1e-6; n_neighbors = 12; algorithm='ball_tree'; leaf_size=48
ROD	contamination = 1e-6;

Appendix 1C: Friedmans test with iman davenports correction

Given a data matrix made up of a specific performance metric (e.g., F0.5score) with N rows representing scenarios in LeakDB benchmark, where $1 \leq i \leq N$, and M columns representing anomaly detection methods, where $1 \leq j \leq M$, the procedure for the Friedman's test is outlined as follows.

1 For each scenario i, let r_i^j denote the rank of the j-th AD method. r_i^j takes values from 1 (best performance) to M (worst performance). When there are ties, assign average ranks.
2 For each algorithm j, average the ranks across all scenarios to obtain its average rank denoted R_j.

$$\mathrm{R_j} = \frac{1}{N}\sum_{i=1}^{N} r_i^j \tag{4}$$

3 Compute the Friedman's statistic and the Iman Davenport's correction as follows.

$$\chi_F^2 = \frac{12N}{M(M+1)}\left[\sum_j R_j^2 - \frac{M(M+1)^2}{4}\right] \tag{5}$$

$$F_{ID} = \frac{(N-1)\chi_F^2}{N(M-1)-\chi_F^2}$$

Where the Friedman statistic, χ_F^2, is distributed according to a χ^2 distribution with $M-1$ degrees of freedom when $N > 10$ and $M > 5$. The Iman Davenport statistic, F_{ID}, is distributed according to the F-distribution with $M-1$ and $(M-1)(N-1)$ degrees of freedom.

Appendix 1D: TOPSIS MCDM Framework

Given alternatives (semi-supervised AD methods) and criteria ($F_{0.5}$Score, PR AUC Score, and Identification Tine Lag), the algorithm for the TOPSIS MCDM framework utilized in this study is presented in the following 7 steps:

Step 1: Construct the decision matrix

The Decision Matrix (DM) is formulated such that each column represents a criterion to be utilized in the evaluation of the semi-supervised AD methods. The DM is given as:

$$\mathrm{DM} = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_n \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_m \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \end{array} \tag{6}$$

Where $\mathrm{A_i},\ i = 1,\ 2, \ldots, \mathrm{n}$ represents the alternatives to be evaluated and ranked. $\mathrm{C_j} = 1,\ 2, \ldots, \mathrm{m}$ denotes the criteria to be utilized for the evaluation. Each element of the DM denoted $\mathrm{x_{ij}}$ represents an alternative i with respect to criterion j.

Step 2: Normalize the Decision Matrix

The Decision Matrix is normalized to compute the relative performance of each alternative considering all other alternatives for each criterion.

Each element in the normalized decision matrix, R_{ij} is mathematically expressed as

$$R_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}} \tag{7}$$

Where n is the total number of alternatives to be considered.

Step 3: Compute the weighted Decision Matrix

In order to compute the weights signifying the relative importance of each criterion objectively, two weighting approaches, equal weights, and entropy weights, are implemented.

a) **Equal Weights**

The equal weight vector w_j, $j = 1, 2, \cdots, m$, which assigns equal importance to each criterion is computed as follows

$$w_j = \frac{1}{m} \tag{8}$$

a) **Entropy Weights**

The entropy weight vector, w_j, $j = 1, 2, \cdots, m$, is computed using the following procedure.

(i) Compute the entropy of each criterion.

The entropy value of each criterion h_j is computed as

$$h_j = -h_0 \sum_{i=1}^{n} p_{ij} \ln(p_{ij}),\ j = 1, 2, \cdots m \tag{9a}$$

where h_0 is the entropy constant and is equal to $(\ln n)^{-1}$, and

$p_{ij} \ln(p_{ij}) = 0$ if $p_{ij} = 0$.

(ii) Compute the degree of diversification.

The degree of diversification denoted d_j is given as

$$d_j = 1 - h_j,\ j = 1, 2, \cdots m \tag{9b}$$

(iii) Compute the objective weight vector.

The weight vector which signifies the relative importance of each criterion is given as

$$w_j = \frac{d_j}{\sum_{k=1}^{m} d_k},\ j = 1, 2, \cdots, m \tag{9c}$$

Finally, the weighted decision matrix is then computed by multiplying each column of the normalized decision matrix by its corresponding weight from one of the weighting approaches.

$$V_{ij} = w_j \times R_{ij} \tag{10}$$

Step 4: Compute the Positive and Negative Ideal Solution

The Positive Ideal Solution (A^+) and the Negative Ideal Solution (A^-) solutions are computed according to the weighted decision matrix via the following equations

$$\text{PIS} = A^+ = \left\{V_1^+,\ V_2^+,\ \cdots, V_m^+\right\} \text{ where } V_j^+ = \left\{\left[\max_i (V_{ij}) | \ j \in J\right], \left[\min_i (V_{ij}) | \ j \in J'\right]\right\} \tag{11}$$

$$\text{NIS} = A^- = \left\{V_1^-,\ V_2^-,\ \cdots, V_m^-\right\} \text{ where } V_j^- = \left\{\left[\min_i (V_{ij}) | \ j \in J\right], \left[\max_i (V_{ij}) | \ j \in J'\right]\right\} \tag{12}$$

Where J is associated with the beneficial criterion and J' is associated with the non-beneficial criterion.

Step 5: Calculate the separation form the Ideal and Non-Ideal Solution

The separation distance of each alternative from the ideal and non-ideal solution is computed as

$$S_i^+ = \left[\sum_{j=1}^{n_c}\left(V_j^+ - V_{ij}\right)^2\right]^{1/2} \tag{13}$$

$$S_i^- = \left[\sum_{j=1}^{n_c}\left(V_j^- - V_{ij}\right)^2\right]^{1/2} \tag{14}$$

Where, i = alternative index and j = criterion index.

Step 6: Measure the relative closeness of each alternative to the ideal solution.

For each alternative the relative closeness to the ideal solution is computed as.

$$C_i = \frac{S_i^-}{S_i^+ + S_i^-},\ 0 \le C_i \le 1 \tag{15}$$

Step 7: Rank the preference order.

The semi-supervised AD methods are ranked based on the value of C_i, the higher the value of C_i, the higher the ranking order and hence the better the semi-supervised AD method in leakage identification. The overall best performing semi-supervised AD method is the one with the largest value of C_i.

References

Almardeny, Y., Boujnah, N., & Cleary, F. (2020). A novel outlier detection method for multivariate data. *IEEE Transactions on Knowledge and Data Engineering*.

AWWA, W. A. (2016). *Loss control programs (M36): Awwa manual of practice*. Denver: American Waterworks Association.

Ayadi, A., Ghorbel, O., BenSalah, M., & Abid, M. (2019). Kernelized technique for outliers detection to monitoring water pipeline based on WSNs. *Computer Networks, 150*, 179–189.

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Security and Applications, 3*(10).

Benítez, P., Rocha, E., Varum, H., & Rodrigues, F. (2020). A dynamic multi-criteria decision-making model for the maintenance planning of reinforced concrete structures. *Journal of Building Engineering, 27*, Article 100971.

Bergmann, B., & Hommel, G. (1988). *Multiple hypothesenprüfung/multiple hypotheses testing* (pp. 100–115). Springer.

Besner, M.-C., Prévost, M., & Regli, S. (2011). Assessing the public health risk of microbial intrusion events in distribution systems: Conceptual model, available data, and challenges. *Water Research, 45*(3), 961–979.

Boafo-Mensah, G., Neba, F. A., Tornyeviadzi, H. M., Seidu, R., Darkwa, K. M., & Kemausuor, F. (2021). Modelling the performance potential of forced and natural-draft biomass cookstoves using a hybrid Entropy-TOPSIS approach. *Biomass and Bioenergy, 150*, Article 106106.

Breunig, M.M., Kriegel, H.-P., Ng, R.T. and Sander, J. 2000 LOF: Identifying density-based local outliers, pp. 93–104.

Calvo, B., & Santafé Rodrigo, G. (2016). scmamp: Statistical comparison of multiple algorithms in multiple problems. *The R Journal. Vol. 8/1, Aug. 2016*.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys, 41*(3). Article 15.

Cody, R. A., & Narasimhan, S. (2020). A field implementation of linear prediction for leak-monitoring in water distribution networks. *Advanced Engineering Informatics, 45*, Article 101103.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research, 7*, 1–30.

Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation, 1*(1), 3–18.

EurEau. (2017). *Europe's water in figures: An overview of the european drinking water and waste water sectors*. The European Federation of National Water Associations Brussels.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(4), 463–484.

Goldstein, M., & Dengel, A. (2012). Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: Poster and demo track, 9*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research, 13*(1), 723–773.

Haibo, H., & Yunqian, M. (2013). *Imbalanced learning: Foundations, algorithms, and applications, 1* p. 27). Wiley-IEEE Press.

Hashim, H., Ryan, P., & Clifford, E. (2020). A statistically based fault detection and diagnosis approach for non-residential building water distribution systems. *Advanced Engineering Informatics, 46*, Article 101187.

Hu, Z., Chen, B., Chen, W., Tan, D., & Shen, D. (2021). Review of model-based and data-driven approaches for leak detection and location in water distribution systems. *Water Supply, 21*(7), 3282–3306.

Hwang, C.-L., & Yoon, K. (1981). *Multiple attribute decision making* (pp. 58–191). Springer.

Iman, R. L., & Davenport, J. M. (1980). Approximations of the critical region of the fbietkan statistic. *Communications in Statistics-Theory and Methods, 9*(6), 571–595.

Jenks, B. W., & Papa, F. (2022). Mobile DMA unit provides a proactive leakage management strategy. *Opflow, 48*(3), 24–27.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data, 6*(1), 1–54.

Kara, S., Karadirek, I. E., Muhammetoglu, A., & Muhammetoglu, H. (2016). Hydraulic modeling of a water distribution network in a tourism area with highly varying characteristics. *Procedia Engineering, 162*, 521–529.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence, 5*(4), 221–232.

Kriegel, H.-P., Schubert, M., & Zimek, A. (2008). *Angle-based outlier detection in high-dimensional data* (pp. 444–452).

Latecki, L. J., Lazarevic, A., & Pokrajac, D. (2007). *Outlier detection with kernel density functions* (pp. 61–75). Springer.

Li, S., & Li, J. Z. (2009). Hybridising human judgment, AHP, simulation and a fuzzy expert system for strategy formulation under uncertainty. *Expert Systems with Applications, 36*(3), 5557–5564.

Li, Z., Zhao, Y., Botta, N., Ionescu, C., & Hu, X. (2020). *COPOD: Copula-based outlier detection* (pp. 1118–1123). IEEE.

Liemberger, R., & Wyatt, A. (2019). Quantifying the global non-revenue water problem. *Water Supply, 19*(3), 831–837.

Mahalanobis, P. C. (1936). *On the generalized distance in statistics*. National Institute of Science of India.

Mamo, T. G., Juran, I., & Shahrour, I. (2014). Virtual DMA municipal water supply pipeline leak detection and classification using advance pattern recognizer multi-class SVM. *Journal of Pattern Recognition Research, 9*(1), 25–42.

Munaga, H., & Jarugumalli, V. (2011). *Performance evaluation: Ball-treeand kd-tree in the context of mst* (pp. 225–228). Springer.

Muniz Do Nascimento, W., & Gomes-Jr, L. (2022). Enabling low-cost automatic water leakage detection: A semi-supervised, autoML-based approach. *Urban Water Journal*, 1–11.

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(1), 86–97.

Nam, K., Ifaei, P., Heo, S., Rhee, G., Lee, S., & Yoo, C. (2019). An efficient burst detection and isolation monitoring system for water distribution networks using multivariate statistical techniques. *Sustainability, 11*(10), 2970.

Perelman, L. S., Abbas, W., Koutsoukos, X., & Amin, S. (2016). Sensor placement for fault location identification in water networks: A minimum test cover approach. *Automatica, 72*, 166–176.

Ramaswamy, S., Rastogi, R., & Shim, K. (2000). *Efficient algorithms for mining outliers from large data sets* (pp. 427–438).
Roshan, K. A., Tang, Z., & Guan, W. (2019). High fidelity moving Z-score based controlled breakdown fabrication of solid-state nanopore. *Nanotechnology, 30*(9), Article 095502.
Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics : A Journal of Statistics for the Physical, Chemical, and Engineering Sciences, 41*(3), 212–223.
Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one, 10*(3), Article e0118432.
Santos-Ruiz, I.d.l., López-Estrada, F. R., Puig, V., Pérez-Pérez, E., Mina-Antonio, J., & Valencia-Palomo, G. (2018). Diagnosis of fluid leaks in pipelines using dynamic PCA. *IFAC-PapersOnLine, 51*(24), 373–380.
Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation, 13*(7), 1443–1471.
Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., & Chang, L. (2003). *A novel anomaly detection scheme based on principal component classifier*. Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering.
Steffelbauer, D. B., & Fuchs-Hanusch, D. (2016). Efficient sensor placement for leak localization considering uncertainties. *Water resources management, 30*(14), 5517–5533.
Terrell, G. R., & Scott, D. W. (1992). Variable kernel density estimation. *The Annals of Statistics*, 1236–1265.
Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.
Thienen, P., & Vertommen, I. (2016). Automated feature recognition in CFPD analyses of DMA or supply area flow data. *Journal of Hydroinformatics, 18*(3), 514–530.
Tornyeviadzi, H. M., Neba, F. A., Mohammed, H., & Seidu, R. (2021). Nodal vulnerability assessment of water distribution networks: An integrated Fuzzy AHP-TOPSIS approach. *International Journal of Critical Infrastructure Protection, 34*, Article 100434.
Vercruyssen, V., Meert, W., Verbruggen, G., Maes, K., Baumer, R., & Davis, J. (2018). *Semi-supervised anomaly detection with an application to water analytics* (pp. 527–536). IEEE.
Villa-Pérez, M. E., Alvarez-Carmona, M. A., Loyola-González, O., Medina-Pérez, M. A., Velazco-Rossell, J. C., & Choo, K.-K. R. (2021). Semi-supervised anomaly detection algorithms: A comparative summary and future research directions. *Knowledge-Based Systems, 218*, Article 106878.
Vrachimis, S.G. and Kyriakou, M.S. 2018 LeakDB: A benchmark dataset for leakage diagnosis in water distribution networks.
Wu, Y., & Liu, S. (2017). A review of data-driven approaches for burst detection in water distribution systems. *Urban Water Journal, 14*(9), 972–983.
Zhao, Y., Nasrullah, Z. and Li, Z. 2019. Pyod: A python toolbox for scalable outlier detection. *arXiv preprint arXiv:1901.01588*.