Pattern Detection in Abnormal District Heating Data

Gideon Mbiydzenyuy(✉) and Håkan Sundell

University of Boras, 501 90 Boras, Sweden
{gideon.mbiydzenyuy,hakan.sundell}@hb.se
https://www.hb.se/en/

Abstract. Data generated by industrial systems such as District Heating (DH) often lack meaningful labels for supervised Machine Learning (ML) methods in anomaly detection. Consequently, unsupervised and semi-supervised ML methods are widely used. These methods frequently uncover numerous anomalies, necessitating labor-intensive post-processing. This paper proposes an algorithm to detect topK anomaly instances with similar patterns (energy signatures) to known anomalies, and to identify clusters of similar anomalies using hierarchical clustering. Similarities between anomaly instances are computed using Dynamic Time Warping and Matrix Profiles. Generative Adversarial Networks (GANs) are employed to augment small anomaly datasets. Results demonstrate the effectiveness of the proposed algorithm in reducing the manual effort required for post-processing anomalies in a DH dataset.

Keywords: anomaly pattern · cluster · hierarchical · similarity · labels

1 Introduction

Anomaly detection plays a crucial role in reducing maintenance costs for many industrial systems [5,22,27,30]. Unlike supervised Machine Learning (sML), which relies on labeled datasets defining normal and abnormal conditions [30], many industrial datasets, such as those from District Heat (DH) networks, do not have suitable labels for sML approaches [17,22]. Instead, unsupervised ML (uML) and semi-supervised ML are widely adopted for anomaly detection in industrial systems [23,26,29].

Each node in a DH network consists of a building, which can be multi-dwelling, private, or industrial (e.g., hospitals and schools), each with different energy profiles. For DH networks, data often comprises multivariate time series detailing parameters such as supply temperature, return temperature, flow rate,

The work presented in this article is financed by the Swedish Knowledge Foundation (KKS http://www.kks.se/om-oss/in-english/) under grant no Dnr.20170182 within the project Data Analytics for Fault Detection in District Heating (DAD).

G. Nicosia et al. (Eds.): LOD 2024/ACAIN 2024, LNCS 15508, pp. 224–239, 2025.
https://doi.org/10.1007/978-3-031-82481-4_16

and pressure, primarily recorded for invoicing energy consumption at customer substations [8,19]. To improve DH network efficiency, engineers must continuously monitor, detect, and interpret abnormal sensor signatures [12]. Detection of such anomalies triggers alarms, often with a high rate of false negatives [28]. Effective analysis of signatures across a larger network requires dedicated techniques to complement current anomaly detection systems [1].

Time series datasets pose challenges due to their simultaneous consideration of spatial and temporal characteristics, compounded by the fact that same phenomena, e.g., pressure loss and leakages, may not always result in a unique time series profile and therefore lack labels [1,27]. Therefore, uML methods such as Gaussian mixture model (GMM), Ordering Points To Identify Cluster Structure (OPTICS), Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [11], Autoencoders, and One-Class Support Vector Machines (OCSVM) are commonly employed [3,9,29]. These methods excel in anomaly detection without requiring explicit labels but provide limited insights into the nature of anomalies.

In time series analysis, uML methods can identify various types of anomalies [11,24]: 1) point-based anomalies, where a single data point or sequence deviates abruptly from the norm, often caused by sudden spikes; 2) contextual anomalies, which depend on external conditions such as increased energy consumption in specific parts of the DH network due to weather variations; and 3) collective anomalies, where a group of data points collectively form an anomaly, even though individual points may not be anomalous. For instance, anomalies in energy consumption across different seasons may exhibit different patterns among buildings within a DH network. Interpreting these anomalies is challenging, particularly because instances may belong to multiple categories or types, and distinguishing incipient faults from normal variations or background noise is complex.

This study focuses on identifying similarities among contextual anomalies in order to enhance anomaly interpretation and management in DH networks. It diverges from existing literature [1,23,26,30], which predominantly focuses on applying sML and uML for anomaly detection in industrial datasets. Instead, we propose a semi-supervised approach specifically designed to post-process anomalies identified by uML in DH networks. This method aims to enhance the interpretation and actions in industrial systems based on insights derived from uML-detected anomalies, crucial for implementing timely intervention strategies such as early maintenance of DH substations. Unlike approaches that deal with datasets containing both anomalies and normal data [2,23], our motivation stems from the frequent occurrence of large numbers of anomalies detected by uML methods. Without effective techniques for post-processing such anomalies, engineers default to labor-intensive and time-consuming analyses [28]. This not only incurs significant costs but also hampers the ability to implement timely maintenance intervention strategies.

By scrutinizing patterns within sets of collective and contextual anomalies, our approach seeks to streamline the analysis process, reducing the labor-intensive nature of anomaly analyses. Ultimately, this could significantly mitigate maintenance costs associated with industrial networks by enabling proactive strategies to address anomalies promptly and effectively.

1.1 Problem Statement

The problem consist of finding a suitable method for post-processing a large set anomalies resulting from uML methods when used in the analysis of time series data (univariate and multivariate). Consider the following Eq. (1).

$$\mathbf{A_{n\times\tau}} = \overbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1\tau} \\ x_{21} & x_{22} & \cdots & x_{2\tau} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{n\tau} \end{bmatrix}}^{time} \left.\vphantom{\begin{matrix} x \\ x \\ x \end{matrix}}\right\}\text{instances} \tag{1}$$

Symbol	**Description**
X	Multivariate dataset from a District Heating (DH) network
$A \subset X$	Subset of anomalies generated by an unsupervised Machine Learning (uML) method, e.g., OPTICS, PCA
$n = \|A\|$	Number of anomaly instances resulting from uML applied to X
τ	Time period, e.g., winter season
$x_{it}, i \in 1..n, t \in \tau$	Known anomaly instance indicated by integer i with values recorded over time $t \in \tau$

To address this, we aim to solve the following two problems:

1. **Problem I**: Assuming we know the cause and pattern of a specific anomaly instance $x_{it}, \in A, i \in 1..n, t \in \tau$, find a method to detect similar anomalies $\hat{x}_{it}, i \in 1..n, t \in \tau$ in a given anomaly dataset A.
2. **Problem II**: Assuming we do not have any information about any specific anomaly in A, we aim for a suitable method to explore the presence of clusters of anomalies in A that share similar characteristics. One way to achieve this is to generate p anomaly clusters where cluster k is represented by $C_k \subset A$, $k = 1, 2, ..., p$. Members of each cluster C_k are anomaly instances that share a similar pattern or profile and are distinct from members of other clusters.

Similarity is constructed through a profile comparison among anomaly instances over a specified time window. The result could be zero or multiple anomalies with similar profiles. In the case of multiple anomalies, the result can be ranked to emphasize the most similar anomalies to the given instance.

1.2 Problem Background

Unsupervised Machine Learning (uML) methods for anomaly detection in multivariate time series data are employed to detect anomalies A in industrial dataset X by identifying data points that deviate significantly from the majority. This raises a fundamental question: can an anomaly dataset such as A reveal internal similarities among its instances?

Most uML methods (e.g., OPTICS, PCA) primarily focus on the distribution of the normal data in industrial dataset, identifying patterns or clusters that represent such data. Anomalies are typically data points that do not conform to any established cluster. In domains such as DH, uML methods generate numerous anomalies characterized by complex mathematical properties, stemming from their distinct dissimilarity to normal data. This complexity underscores the need for innovative techniques to effectively analyze these anomalies. Detecting and locating process-specific patterns in industrial time series data is a common task done by engineers [12]. Discovering similarities among anomalies not only facilitates proactive maintenance strategies but also enables the development of labeled datasets for further analysis and refinement of anomaly detection models.

Fitting models to anomaly datasets is challenging due to the need to understand their unique characteristics in order to select suitable ML techniques [9,29]. In the context of DH networks the following observations can be made; 1) operational conditions of DH system sensors often exhibit similarities, such as outdoor temperatures, usage patterns, and intrinsic sensor parameters. Consequently, similarities in faulty or abnormal signatures, where operational conditions play a significant role, are plausible. For instance, abnormal signatures may result from reduced sensitivity of control valves in DH substations albeit to varying extent. 2) Since time series data capture temporal and spatial conditions, they enable comparison of evolutionary signatures over time and space, accounting for phenomena like non-stationarity (concept drift) and phase shifts in frequency and time domains. These systematic changes underscore the dynamic nature of time series data. 3) In industrial applications like DH typically deploy a limited variety of sensor types within specific networks, so sensors installed within the same period often exhibit similar signatures in both normal and degraded operations. Additionally, similarities in abnormal signatures may arise from consistent operational practices by the same group of engineers over extended periods. The above observations gives us reason to believe that anomaly dataset from DH could exhibit reasonable similarities that can be detected in order to facilitate post-processing of anomalies.

This work is pioneering in focusing on post-analysis of anomalies as a separate issue from anomaly detection. Previous studies combined post-analysis with detection [12,26], giving limited attention to post-analysis, especially in categorical [4,20] or labeled data contexts [12]. Additionally, this article demonstrates tailoring GAN networks to address label scarcity in anomaly detection. This paper proposes an algorithm to detect similarities in anomaly datasets. Section 2 presents related work, while Sect. 3 presents the proposed method, with experiment setup and results and conclusions in Sect. 4, 5 and 6, respectively.

2 Related Work

Anomaly detection in industrial datasets, particularly in district heating (DH) networks, has evolved significantly, driven by several themes in existing research [1,12,17–19,22,24–26]. Initially, the focus was on developing unsupervised machine learning (uML) methods to detect anomalies in multivariate time series data without heavy reliance on labeled datasets. Techniques such as Gaussian Mixture Models (GMM), OPTICS, and Autoencoders have been prominent [26,33]. Anomaly detection in DH networks is particularly complex due to diverse temporal and spatial features [1,2]. Advances in computational power have enabled the development of sophisticated similarity measures, facilitating refined comparisons among anomalies [6,16]. Moreover, the availability of large, unlabeled datasets coupled with generative models has leveraged abundant data to enhance model robustness [7]. Another significant trend has been the shift from traditional anomaly detection to fault detection and diagnostics [22,23], reflecting a broader industry movement towards deriving actionable insights. This transition underscores the growing importance of interpreting anomaly patterns to optimize maintenance strategies and reduce operational costs in industrial contexts [22].

Recent advancements demonstrate a significant shift from traditional anomaly detection methods towards specialized fault detection techniques across various domains. Ramírez-Sanz [23] emphasize fault diagnosis using machine learning algorithms tailored for specific faults in complex systems, integrating domain knowledge with data-driven approaches. Mercorelli [21] provide a comprehensive overview of fault diagnosis techniques in industrial processes, highlighting challenges related to the practical applications of detection methods. These studies collectively illustrate a paradigm shift towards sophisticated fault detection approaches, leveraging advanced data analytics and machine learning to ensure early detection and mitigation of faults in dynamic systems.

Pattern detection in time series data has attracted interest in related domains [33], where various anomaly patterns such as Novelty Pattern, Surprise Pattern, Discord, Novel Event, and Aberrant Behavior across different domains have been reported. Yu et al. [33] proposed algorithms using window-based, similarity-based, symbolic representation-based, and model-based techniques for anomaly detection, employing Trend Feature Symbolic Aggregate approximation (TFSAX) and weighted Probability Suffix Tree (wPST) for sequence analysis. Their approach identifies top-k pattern anomalies effectively. Deng et al. [6] introduced a white-box machine learning approach for multivariate time series, enhancing Dynamic Time Warping (DTW) for efficient similarity calculation and pattern extraction. They used sliding window strategies for local anomalies and cumulative window analysis for long-term anomalies, employing fast-DTW and Euclidean distance for similarity scoring. Luczak et al. [16] explored

Derivative Dynamic Time Warping (DDTW) to improve hierarchical clustering of time series data, emphasizing its application in capturing variability for anomaly detection. These studies highlight diverse methodologies for similarity measurement in time series analysis, tailored to specific data characteristics and application requirements.

Generative Adversarial Networks (GANs) play a pivotal role in augmenting data for training and testing pattern detection methods. Esteban et al. [7] introduced Recurrent GANs (RGANs) and Recurrent Conditional GANs (RCGANs) for generating realistic multi-dimensional time series data, crucial in scenarios with limited labeled data. Li et al. [14] developed DoppelGANger (DG), enhancing fidelity of synthetic time series data across various domains, including cybersecurity and healthcare. Leveraging GAN-augmented datasets enhances robustness and generalization capabilities of pattern detection algorithms, ensuring effective anomaly detection in complex, dynamic environments.

Anomaly detection methods often yield extensive lists of anomalies, posing challenges in analysis and interpretation across domains. Yu et al. [33] and Deng et al. [6] highlight this issue, utilizing window-based algorithms and DTW for anomaly detection. However, the sheer volume of anomalies necessitates efficient methods for meaningful analysis. Pattern detection offers promise, as Das et al. [4] emphasize the identification of recurring patterns within anomalies, enhancing interpretability and providing deeper insights. Existing methodologies are often domain-specific, with limited application in district heating. Addressing this gap, this study introduces an integrated framework using pattern detection tailored for district heating systems, aiming to enhance anomaly detection effectiveness and enabling proactive anomaly management strategies.

3 A Method for Anomaly Pattern Detection

The Anomaly Pattern Detection Algorithm 1 is designed to analyze an anomaly dataset A (1) using various similarity measures. Algorithm 1 either returns 1) *topK*-similar instances to a given anomaly instance, 2) *topK*-similar instances to all anomaly instances and, 3) hierarchical clusters of instances for all anomaly instances in A. We design and train a Generative Adversarial Netork (GAN) with the purpose of augmenting the anomaly dataset A to evaluate the proposed Algorithm 1. The proposed Algorithm 1 proceed in the following phases:

- Given an anomaly dataset A Eq. 1, a labeled anomaly instance k, a window size ws, and an integer value $topK$.
- Initialize an empty list V to track visited nodes and another list S_k to store similarity scores of visited nodes for the given instance k. Anomaly dataset A is transformed using the relative change calculation Eq. 2.
- Compute pairwise similarity matrix (symmetric) with distances based on Derivative Dynamic Time Warping (DDTW) or Matrix Profile (MPDist) with a rolling window ws for instances in transformed A.
- For each instance $i \in A$, compute and sorts $topK$ similarity scores S_k with instance k.

- Iterates through pairs of anomaly instances (i, j) in transformed A and sort *topK* similarity scores with an arbitrary instance j forming clusters $C_k \subset A$, $k = 1, 2, ..., p$.
- Hierarchically form clusters $hc(S)$ on pairwise similarity matrix of transformed A.
- Return *topK*, $C_k \subset A$, $k = 1, 2, ..., p$ and $hc(S)$.

Algorithm 1. Anomaly Pattern Detection Algorithm

```
procedure AnomalyPatternDetection($A, k, ws, topK$)
    $V \leftarrow [\ \ ], S_k \leftarrow [\ \ ]$                ▷ Visited nodes and similarity scores
    $S \leftarrow [\ \ ]$                                    ▷ $n \times n$ similarity matrix
    for $i, j \in \binom{n}{2}$ do
        $x_{it} \leftarrow x'_{it}, x_{jt} \leftarrow x'_{jt}$           ▷ Compute relative change
        $S[i,j] \leftarrow \begin{cases} \text{DDTW}(A[i], A[j], ws) & \text{if DDTW is used} \\ \text{MPDist}(A[i], A[j], ws) & \text{if MPDist is used} \end{cases}$
    for $i \in A$ do
        $S_k[i] \leftarrow S[i,k]$      ▷ Given $k$ compute topK similar anomalies (Problem I)
        if $S_k[i] \neq \emptyset$ then
            $S_k[i] \leftarrow \text{Sort}(S[i,k])_{topK}$
    for $(i, j) \in A, i \neq j$ do
        if $i \notin V$ then
            if index($i$) == 0 then
                $S_j \leftarrow S[i,j]$    ▷ Compute topK of any instance $j \in A$, Problem (II)
                if $S_j \neq \emptyset$ then
                    $S_j \leftarrow \text{Sort}(S[i,j])_{topK}$
                    $V \leftarrow j$
            else
                $A \leftarrow A \setminus S_j$
                $S_j \leftarrow S[i,j]$
                if $S[i,j] \neq \emptyset$ then
                    $S_j \leftarrow |S[i,j]|_{topK}$
            $V \leftarrow j$
            $A \leftarrow A \setminus S_j$
    $C \leftarrow \text{hc}(S)$       ▷ Compute hierarchical clusters for all instances (Problem II)
    return $C, S_k$
```

In our study, we rigorously evaluate the quality of our proposed anomaly detection algorithm by leveraging a combination of hierarchical clustering techniques and advanced statistical metrics. The effectiveness of our clustering approach is quantified using the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index similar to the external measures in [16] which together assess the compactness, separation, and similarity of the clusters. Visualizations of cluster formations further validate the integrity and practical utility of the detected anomalies. The optimal number of clusters is determined through an iterative process employing, which calculates linkage from precomputed pairwise distances and evaluates various cluster configurations to pinpoint the most effective clustering strategy.

This approach is structured to identify patterns and similarities within the anomaly dataset A, leveraging different similarity metrics and clustering techniques to facilitate anomaly pattern detection and analysis. We evaluate the quality of generated anomalies using the remaining section presents the details of the computations in Algorithm 1.

3.1 Distance Measures for Computing Similarities

A number of distance measures can be employed in the proposed algorithm when computing similarity between any pair of time series anomaly instances representing DH energy profiles.

Relative Change: To compute the similarity score $S_{i,j}$ between two instances $x_{i,t}$ and $x_{j,t}$ (where $i, j \in 1..n, t \in \tau$), we calculate their relative change between consecutive measurement points using Eq. 2:

$$x'it = \frac{x_{i,t} - x_{i,t-1}}{x_{i,t-1}} \tag{2}$$

Here, when the result is division by zero, the implementation returns 'NAN' which is then filled with the preceding value, ensuring that the relative change computation is well-defined even when $x_{i,t}$ is close to zero. This adjustment is crucial, especially since missing values are replaced with zeros, as mentioned earlier. This relative change computation over a time window shifts focus from absolute sensor readings to their variations over time. Anomalies often manifest as significant deviations in these relative changes, making the relative differential a suitable measure for capturing temporal variations.

Derivative Dynamic Time Warping (DDTW): Dynamic Time Warping (DTW) is a distance measure used to assess similarities between time series instances $(x_{i,t}, x_{j,t})$. It calculates the cost of aligning two time series by minimizing the cumulative distance between corresponding points, allowing for non-linear time distortions. The traditional DTW method compares raw values directly and then uses these to construct a wharping path. In the context of similarities in anomaly instances, Derivative Dynamic Time Warping ($DDTW$) extends DTW by computing the relative change of the time series data over a time window unlike the approach proposed by Mciej [16]. To do this we compute the differences between consecutive values in the time series, as shown in Eq. 2 over a time window. Shorter time windows will amplify local changes in the time series while longer time windows amplify trends over the time window. This approach is particularly effective for finding similarities in anomalies characterized by subtle variations in their temporal patterns (incipient anomalies) as a result of concept drift, which may not be apparent when using traditional DTW alone.

Matrix Profile Distance Measure (MPDist): A matrix profile transforms a time series into an array that records similarity scores computed by varying a sliding window of sub-sequences along the time series [10]. Given two instances x_{it}, x_{jt} where $i, j \in \{1, \ldots, n\}$ and $t \in \tau$, their matrix profile can be computed and sorted in both (i, j) and (j, i) directions to identify similarities in sub-sequences regardless of their start and end points. This is denoted as $\hat{x}_{ij}^{ji}$, where $i, j \in \{1, \ldots, n\}$ and $t \in \tau$, as detailed in [10]. The parameter $\hat{k}$ is set to 5% of $2l$, where l is the length of x_{it}, $i \in \{1, \ldots, n\}$, and $t \in \tau$.

The matrix profile distance measure, $MPDist_{i,j}$, is defined in Eq. 3 based on the sorted similarity structure:

$$MPDist_{i,j} = \begin{cases} \hat{k}^{th}\text{-value of sorted } \hat{x}_{ij}^{ji} & \text{if } |\hat{x}_{ij}^{ji}| > \hat{k} \\ \max(\hat{x}_{ij}^{ji}) & \text{if } |\hat{x}_{ij}^{ji}| \leq \hat{k} \end{cases} \tag{3}$$

This approach is effective for handling abnormal data instances, where smaller $\hat{k}^{th}$-values often result in missing data. The matrix profile provides additional similarity values for sub-sequences beyond Euclidean distance, which it approximates when the sub-sequence length equals the full time series length being analyzed.

Both DDTW and MPDist use smaller values to indicate higher similarity between time series instances. While MPDist offers computational efficiency compared to DDTW, it does not satisfy all properties of a traditional distance metric [10].

3.2 Anomaly Data Augmentation with GAN

Recent studies have shown that GAN networks can potentially be employed to generate useful synthetic time series data [7,13,15]. A novel framework for generating realistic time-series data has been proposed, leveraging the flexibility of unsupervised learning with the control provided by supervised training [32]. The GAN network design incorporates a supervised loss function that captures the conditional distribution within the time series data by using the original data as a reference for training (supervision). Recent studies have demonstrated the potential of GAN networks in generating synthetic time series data that closely mimic real-world distributions [7,13,15]. We configure and train a GAN model based on the article “Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions” [14] and associated open source code implementation.

We selected the maximum length of the sequence as the sample length while the model was fed two weeks sequences $(24 * 7 * 2)$. The sample length choice ensure that the model learned the entire sequence while the two weeks sequences ensures that during training, the model can capture the energy consumption behavior with cycles of two weeks. One feature layer and three attribute layers each with 350 units and a gradient penalty coefficient of 4, batch-size 1024, number of epochs 35000 were chosen after trial an error experimentation. Other configuration parameters included [**Maximum Sequence Length:** τ**, Sample Length:** $0.5 * \tau$**, Attribute Noise Dimension:** $24 * 7 * 2$**, Batch Size: 1024, Number of Epochs: 35000, Gradient Penalty Coefficient: 15, GPU Usage: True**].

4 Experiment

Algorithm 1 was implemented in Python3 using Scipy's standard hierarchical clustering and Dynamic Time Warping (DTW) implementation from [31]. The GAN model was trained on a NVIDIA Tesla V100 GPU with 32 GB GPU memory and an Intel Xeon CPU with 36 GB RAM.

4.1 Dataset

The dataset A comprises hourly anomaly records of heat energy consumption in a mix of apartment buildings and private villas from January to March 2018 ($\tau = 3$ months), delivered via a Swedish DH network in Southern Sweden. While our proposed approach remains agnostic to the specific anomaly detection algorithm used, the anomalies in this study were generated using a density-based method, OPTICS, applied to a dataset encompassing approximately 5000 substations. Additionally, the DH company provided four labeled instances: ***1.) Incorrect settings for the heat control curve in February, 2.) Sensitivity drop of a control valve in January, 3.) Temperature sensor located far from the heat exchanger in March, 4.) Incident involving a hand-operated control valve in February.*** Due to the limited labeled instances, insufficient for a full evaluation of Algorithm 1, sequences were padded to uniform length and fed into a DGAN model configured based on Sect. 3.2 parameters from experimental trials. Evaluation of the DGAN model included comparing Wasserstein and energy distances, with low values (0.008 and 0.015 respectively) indicating strong alignment between synthetic and real datasets. Predictive power was assessed using four models with Synthetic-to-Real (TSTR) and Real-to-Real (TRTR) testing, following Esteban et al. [7].

We conducted experimental trials to select a diverse set of ML algorithms from the Scikit-learn library to evaluate the predictive power of data generated by the GAN model. These includes: 1) Ridge Regressor, 2) Random Forest Regressor (n_estimators = 200, max_depth = 15, min_samples_split = 5), 3) SVR (kernel = 'rbf', C = 10, epsilon = 0.05, gamma = 'scale'), 4) Gradient Boosting Regressor (n_estimators = 300, learning_rate = 0.01, max_depth = 3), and 5) Multilayer Perceptron (MLP) Regressor (hidden_layer_sizes = (150, 75), activation = 'relu', solver = 'adam', alpha = 0.001, max_iter = 500). We aimed at keeping default parameters with minimal fine-tuning. Each experiment used an anomaly dataset split sequentially, training on two months and testing on one month (30% of the dataset). Results for MSE, MAE, RMSE, and R2 Score are shown in Fig. 1. Similar result were obtained when the test set was synthetic.

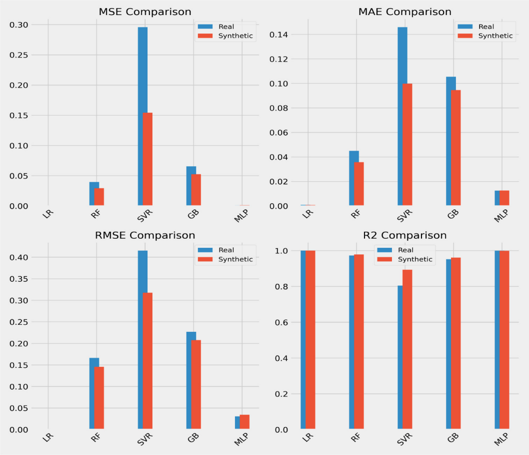

Fig. 1. Comparison of predictive power for different models trained on real and synthetic anomaly data, tested on real data

Figure 1 illustrates that models trained on synthetic data generated by the GAN model and tested on real data performed comparably to models trained and tested on real data, indicating close behavior capture of the generated data. The trained GAN model was used to generate 212 labeled synthetic instances, combined with 212 unlabeled real instances, resulting in a mixed anomaly dataset of 424 instances.

4.2 Algorithm Analysis

For our anomaly detection algorithm, we used a rolling window of 24 h to compute percentage changes (see Eq. 2) and set $ws = 24$ to compute pairwise similarities using DDTW and MPDist. We experimented with various $topK$ values and found $topK = 5$ to be the most effective for identifying meaningful patterns in our dataset. The algorithm's complexity depends on $topK$ pairwise comparisons and the distance matrix computation with DDTW ($O(N^2)$) and MPDist. We addressed this by parallelizing on DGX Nvidia CPU cores (36 out of 40) and using fastDTW implementation [31]. With these adjustments, the computation for the complete dataset took 18 h mainly due to DDTW. Although this limits scalability, once the pairwise distance matrix is computed and saved, adding new instances from the same time frame does not require re-computing the entire dataset. Additionally, using a window-based approach can significantly speed up computations, as a one-month window could be sufficient in practical scenarios. Evaluating Algorithm 1 typically involves benchmark datasets and comparisons with alternative algorithms. However, due to the unique focus of this study on detecting patterns in time series anomaly datasets, traditional benchmarks and methods are not directly applicable. The challenges include: 1) The lack of dedicated methods for anomaly pattern detection in time series datasets, and 2)

Existing benchmark datasets do not combine labeled and unlabeled anomalies, requiring the use of a GAN network to augment the data. Consequently, our evaluation primarily focuses on internal validation of Algorithm 1. [13] comparing the distribution of the synthetic data with the real data instances.

5 Result of Pattern Detection in Anomaly Dataset

5.1 Problem I - topK Detection in Real Data Consisting of Labeled and Unlabeled Anomalies

Given the anomaly dataset from OPTICS and labeled instances from the energy company, Algorithm 1 detected instances by iterating through each labeled anomaly. We experimented with *topK* values ranging from 2 to 50 and evaluated the Silhouette score using precomputed DDTW and MPDist, which were 0.28 and 0.27, respectively when $topK = 5$. The results from both distance matrices for different values of *topK* did not show significant variation. The low scores were due to our focus on extracting individual clusters for each labeled instance, which left the remaining data in a single cluster. Figure 2 presents the result of the top four clusters for each labeled cluster. The cluster of anomalies with 'Hand-operated control valve' indicates that the sensor values took on discrete values in the form of a stepping function. Similar instances with such behavior are visible in the cluster plot, suggesting that the clustering results are logical upon inspection.

Figure 2 The black plot indicate the labeled instance.

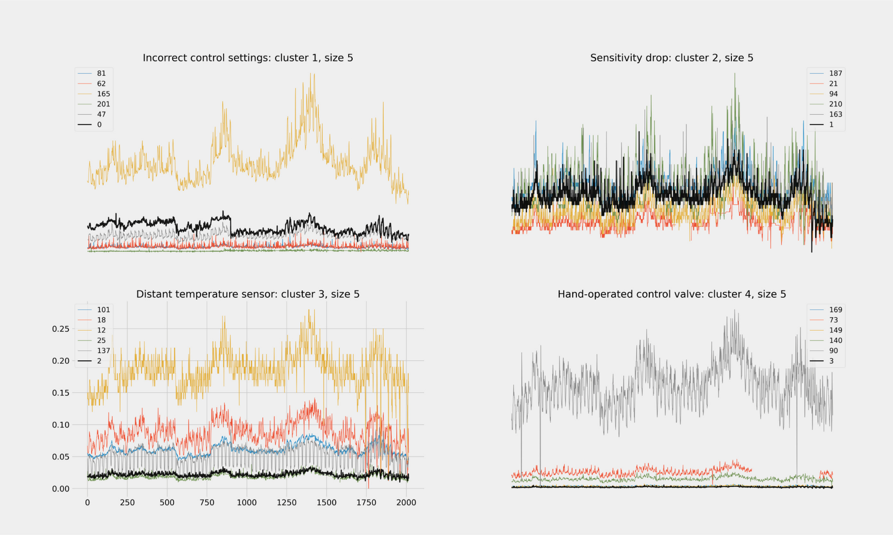

Fig. 2. K = 5 most similar patterns to unlabeled anomalies (MPDist similarity measure).

5.2 Problem II - Cluster Formation with Mixed Real and Synthetic Anomaly Dataset

In this experiment, we evaluated Algorithm 1 using a larger dataset consisting of mixed GAN-generated anomalies and real anomalies. The maximum number of clusters was set to 40, determined through a grid search (2 to 100) based on the maximum Silhouette score, which was 0.63. The hierarchical clustering method organizes the complete DH anomaly dataset into a hierarchical structure of clusters, as shown in Fig. 3. This hierarchical structure facilitates the analysis of anomalies by revealing relationships within and across clusters. Energy analysts can explore clusters starting from the leaves, revealing decreasing similarities within and across clusters.

The algorithm successfully separated synthetic data instances into different clusters compared with DDTW and MPDist distance matrices. Figure 3 illustrates cluster pattern variations at any given hour in a substation (time on the y-axis and substation ID on the x-axis). A drift in the energy profile of substations over time is visible at the peaks for each substation along the x-axis. Based on such a drift, the most inefficient clusters and substations are evident, such as instance 175 in cluster 19 of Fig. 3. This provides a practical starting point for analyzing large anomaly datasets. This provides a practical starting point for analyzing large anomaly datasets. Each cluster can be analyzed in turn and eventually receive a similar label, further improving inputs for supervised ML methods.

5.3 Effect of Distance Metrics, DDTW and MPDist

In the evaluation of hierarchical clustering performance using DDTW and MPDist, MPDist was effective in forming well-defined clusters. The Silhouette Score for MPDist was significantly higher (0.63) compared to DDTW (0.12), indicating more distinct and coherent cluster formations. Additionally, MPDist achieved a higher Calinski-Harabasz Index at 855.26, compared to 196 for DDTW, which supports the finding of better-separated and denser clusters with MPDist. However, the Davies-Bouldin Index, which should ideally be lower, was higher for MPDist (1.3) than for DDTW (0.77), suggesting that while MPDist clusters are distinct, they might not be as compact as those produced by DDTW.

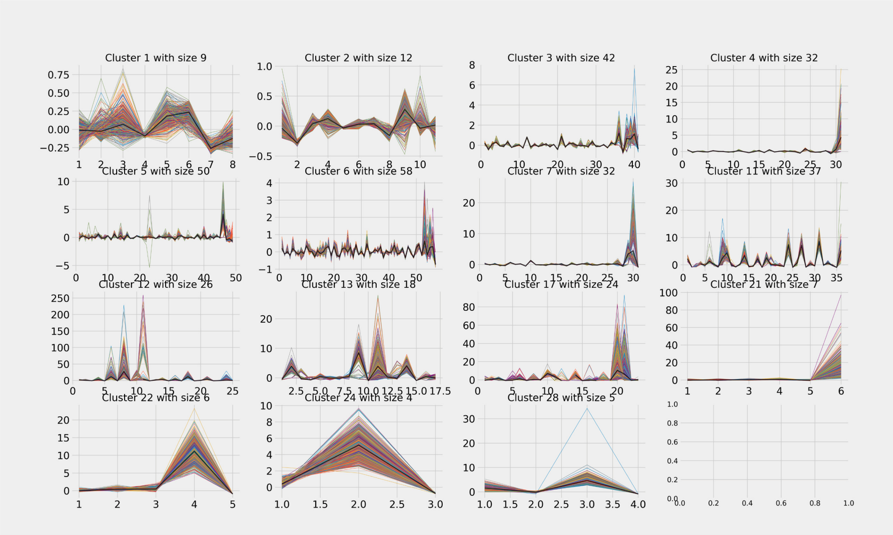

Fig. 3. Substation anomaly patterns.

6 Conclusion

Unsupervised anomaly detection in industrial applications, such as District Heating (DH), often results in numerous anomalies that require intensive manual analysis due to the lack of labels. This article proposes an Anomaly Pattern Detection algorithm to detect topK anomalies with similar patterns, either to a single instance or a set of labeled instances. We employed hierarchical clustering to extract anomaly clusters using similarity measures like matrix profiles and Derivative Dynamic Time Warping (DDTW) on anomaly dataset. Generative Adversarial Networks (GANs) were used to augment the anomaly dataset by generating additional labeled anomalies. The quality of the GAN model was evaluated using multiple machine learning algorithms such as Ridge Regressor, Random Forest Regressor, Gradient Boosting Regressor, and Multilayer Perceptron. The proposed algorithm, which is distance matrix agnostic, was optimized by parallelizing DDTW computations on CPU cores to handle large datasets.

Using the Calinski-Harabasz Score for cluster purity evaluation, we validated the algorithm internally. For hierarchical clustering, the maximum number of clusters was set to 40, determined through a grid search based on the maximum Silhouette score. The topK = 5 most similar anomaly instances to a labeled anomaly were identified using a similar grid search technique.

Experiments with a DH network anomaly dataset showed no distinct preference between matrix profiles and DDTW as distance measures. Due to the high computational cost of DDTW, matrix profiles proved more efficient for analyzing anomaly patterns. The algorithm effectively identified similar anomalies, reducing the burden on engineers managing large anomaly sets in industrial setups like

DH networks. This work is pioneering in focusing on post-analysis of anomalies as a separate issue from anomaly detection. Previous studies combined post-analysis with detection [12,26], giving limited attention to post-analysis, especially in categorical [4,20] or labeled data contexts [12]. Additionally, this article demonstrates tailoring GAN networks to address label scarcity in anomaly detection. Future work could explore various distance matrices and clustering techniques.

References

1. Bahlawan, H., et al.: Detection and identification of faults in a district heating network. Energy Convers. Manag. **266**, 115837 (2022)
2. Blázquez-García, A., Conde, A., Mori, U., Lozano, J.A.: A review on outlier/anomaly detection in time series data. arXiv preprint arXiv:2002.04236 (2020)
3. Choi, K., Yi, J., Park, C., Yoon, S.: Deep learning for anomaly detection in time-series data: review, analysis, and guidelines. IEEE Access **9**, 120043–120065 (2021)
4. Das, K.: Detecting patterns of anomalies. Ph.D. thesis Carnegie Mellon University (2009)
5. De Vita, F., Bruneo, D., Das, S.K.: A semi-supervised Bayesian anomaly detection technique for diagnosing faults in industrial IoT systems. In: 2021 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 31–38. IEEE (2021)
6. Deng, Z., Kang, J., Wang, X.: Multidimensional time series analysis for anomaly pattern detection and interpretation. In: 2024 IEEE 4th International Conference on Power, Electronics and Computer Applications (ICPECA), pp. 1371–1375. IEEE (2024)
7. Esteban, C., Hyland, S.L., Rätsch, G.: Real-valued (medical) time series generation with recurrent conditional GANs. arXiv preprint arXiv:1706.02633 (2017)
8. Gadd, H., Werner, S.: Fault detection in district heating substations. Appl. Energy **157**, 51–59 (2015)
9. Garg, A., Zhang, W., Samaran, J., Savitha, R., Foo, C.S.: An evaluation of anomaly detection and diagnosis in multivariate time series. IEEE Trans. Neural Netw. Learn. Syst. **33**(6), 2508–2517 (2021)
10. Gharghabi, S., Imani, S., Bagnall, A., Darvishzadeh, A., Keogh, E.: Matrix profile XII: Mpdist: a novel time series distance measure to allow data mining in more challenging scenarios. In: 2018 IEEE International Conference on Data Mining (ICDM), pp. 965–970. IEEE (2018)
11. Hurst, W., Montañez, C.A.C., Shone, N.: Time-pattern profiling from smart meter data to detect outliers in energy consumption. IoT **1**(1), 6 (2020)
12. Janka, D., Lenders, F., Wang, S., Cohen, A., Li, N.: Detecting and locating patterns in time series using machine learning. Control. Eng. Pract. **93**, 104169 (2019)
13. Jürgens, M., Scholz, C.: Synthetic time series generation using GANs with application in energy economics. Master's thesis, Chalmers Technical University, Gothenburg (2022)
14. Lin, Z., Jain, A., Wang, C., Fanti, G., Sekar, V.: Generating high-fidelity, synthetic time series datasets with doppelganger. arXiv preprint arXiv:1909.13403 (2019)
15. Lin, Z., Jain, A., Wang, C., Fanti, G., Sekar, V.: Using GANs for sharing networked time series data: challenges, initial promise, and open questions. In: Proceedings of the ACM Internet Measurement Conference, pp. 464–483 (2020)

16. Łuczak, M.: Hierarchical clustering of time series data with parametric derivative dynamic time warping. Expert Syst. Appl. **62**, 116–130 (2016)
17. Månsson, S., Benzi, I.L., Thern, M., Salenbien, R., Sernhed, K., Kallioniemi, P.O.J.: A taxonomy for labeling deviations in district heating customer data. Smart Energy **2**, 100020 (2021)
18. Mbiydzenyuy, G.: Univariate time series anomaly labelling algorithm. In: Machine Learning, Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, 19–23 July 2020, Revised Selected Papers, Part II 6, pp. 586–599. Springer (2020)
19. Mbiydzenyuy, G., Nowaczyk, S., Knutsson, H., Vanhoudt, D., Brage, J., Calikus, E.: Opportunities for machine learning in district heating. Appl. Sci. **11**(13), 6112 (2021)
20. McFowland, E., Speakman, S., Neill, D.B.: Fast generalized subset scan for anomalous pattern detection. J. Mach. Learn. Res. **14**(1), 1533–1561 (2013)
21. Mercorelli, P.: Recent advances in intelligent algorithms for fault detection and diagnosis. Sensors **24**(8), 2656 (2024)
22. Neumayer, M., Stecher, D., Grimm, S., Maier, A., Bücker, D., Schmidt, J.: Fault and anomaly detection in district heating substations: a survey on methodology and data sets. Energy **276**, 127569 (2023)
23. Ramírez-Sanz, J.M., Maestro-Prieto, J.A., Arnaiz-González, Á., Bustillo, A.: Semi-supervised learning for industrial fault detection and diagnosis: a systemic review. ISA Trans. (2023)
24. Schmidl, S., Wenig, P., Papenbrock, T.: Anomaly detection in time series: a comprehensive evaluation. Proc. VLDB Endow. **15**(9), 1779–1797 (2022)
25. Søndergaard, H.A.N., Shaker, H.R., Jørgensen, B.N.: Automated and real-time anomaly indexing for district heating maintenance decision support system. Appl. Therm. Eng. **233**, 120964 (2023)
26. Sun, W., Cheng, D., Peng, W., et al.: Anomaly detection analysis for district heating apartments. J. Appl. Sci. Eng. **21**(1), 33–44 (2018)
27. Usmani, U.A., Happonen, A., Watada, J.: A review of unsupervised machine learning frameworks for anomaly detection in industrial applications. In: Science and Information Conference, pp. 158–189. Springer (2022)
28. Vavilis, S., Egner, A., Petković, M., Zannone, N.: An anomaly analysis framework for database systems. Comput. Secur. **53**, 156–173 (2015)
29. Vercruyssen, V., Meert, W., Verbruggen, G., Maes, K., Baumer, R., Davis, J.: Semi-supervised anomaly detection with an application to water analytics. In: ICDM, vol. 2018, pp. 527–536 (2018)
30. Wang, P., Poovendran, P., Manokaran, K.B.: Fault detection and control in integrated energy system using machine learning. Sustainable Energy Technol. Assess. **47**, 101366 (2021)
31. Wu, R., Keogh, E.J.: FastDTW is approximate and generally slower than the algorithm it approximates. IEEE Trans. Knowl. Data Eng. (2020)
32. Yoon, J., Jarrett, D., Van der Schaar, M.: Time-series generative adversarial networks. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
33. Yu, Y., Wan, D., Zhao, Q., Liu, H.: Detecting pattern anomalies in hydrological time series with weighted probabilistic suffix trees. Water **12**(5), 1464 (2020)