Energy and AI 21 (2025) 100548

Contents lists available at ScienceDirect

Energy and AI

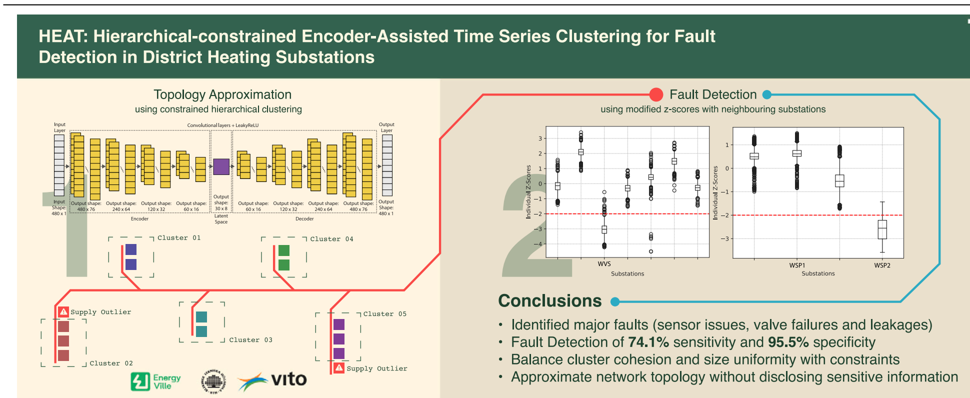
journal homepage: www.elsevier.com/locate/egyai

HEAT: Hierarchical-constrained Encoder-Assisted Time series clustering for fault detection in district heating substations

Jonne van Dreven [a,c,d,*], Abbas Cheddad [a,c], Ahmad Nauman Ghazi [b], Sadi Alawadi [a], Jad Al Koussa [d,e], Dirk Vanhoudt [d,e]

[a] Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden
[b] Department of Software Engineering, Blekinge Institute of Technology, Karlskrona, Sweden
[c] Institute of Computer Science, University of Tartu, Tartu, 51009, Estonia
[d] Unit Energy Technology, Flemish Institute for Technological Research (VITO), Mol, Belgium
[e] EnergyVille, Genk, Belgium

GRAPHICAL ABSTRACT



HIGHLIGHTS

- A novel hierarchical-constrained encoder-assisted time series clustering.
- Integrates soft constraints to enforce cluster sizes and domain-specific information.
- Uses temperatures to approximate a relative topology for localised fault detection.
- Improves performance over conventional clustering and global detection methods.
- Provides an unsupervised and practical solution that enhances operational efficiency.

ARTICLE INFO

Keywords:
Clustering
Semi-supervised learning
Fault detection

ABSTRACT

Fault detection in district heating (DH) substations is crucial for maintaining energy efficiency. However, existing methods often fall short and rely on labelled data or global analysis that may miss subtle anomalies. We introduce HEAT, a Hierarchical-constrained Encoder-Assisted Time series clustering method designed to enhance fault detection in DH substations. HEAT operates in a two-phase approach: first, it approximates a

* Corresponding author at: Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden.
E-mail address: jonne.van.dreven@bth.se (J. van Dreven).

https://doi.org/10.1016/j.egyai.2025.100548
Received 17 March 2025; Received in revised form 6 June 2025; Accepted 30 June 2025
Available online 11 July 2025

District heating

relative network topology using a constraint hierarchical clustering algorithm on supply temperature profiles. HEAT incorporates a Convolutional AutoEncoder (CAE) for dimensionality reduction of the time series data and uses adaptive soft constraints in the linkage function, enabling both minimum and maximum cluster size constraints while supporting domain knowledge, e.g., must-link and cannot-link constraints, using a constraint matrix. Second, we use the topology approximation to perform intra-cluster analysis using Mean Absolute Deviation (MAD) z-scores, with neighbouring substations serving as a validation mechanism, allowing for robust analysis without requiring labelled data. Experimental results demonstrate that HEAT outperforms conventional clustering methods while achieving 74.1% sensitivity and 95.5% specificity in fault detection, significantly improving over typical global analysis. HEAT not only identified major faults (e.g., sensor issues, valve failures) but also detected subtle anomalies (e.g., secondary leakages) while minimising false positives. This unsupervised method offers a viable and flexible solution for DH networks, improving operational efficiency and energy sustainability without disclosing sensitive information.

Abbreviations

4GDH	Fourth-Generation District Heating
AE	Autoencoder
CAE	Convolutional Autoencoder
CI	Confidence Interval
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DENCLUE	Density-based Clustering 2.0
DH	District Heating
DTW	Dynamic Time Warping
FDD	Fault Detection and Diagnosis
HAC	Hierarchical Agglomerative Clustering
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HEAT	Hierarchical-constrained Encoder-Assisted Time series clustering
HHC	High Heat Curve
IEA	International Energy Agency
kM	k-Means Clustering
kS	k-Shape Clustering
LSL	Large Secondary Leakage
LV	Large Valve
MAD	Median Absolute Deviation
SC	Spectral Clustering
WSP	Wrong Sensor Placemen

1. Introduction

Urban heating constitutes a substantial share of global energy consumption, accounting for approximately 50% of the total final energy use, exceeding electricity (20%) and transport (30%). Additionally, it is responsible for over 40% of energy-related carbon dioxide emissions worldwide [1], posing challenges to climate change mitigation, environmental sustainability, and economic resilience [2]. As urban populations expand, the demand for efficient and sustainable heating solutions increases. District heating (DH) offers a promising solution to this problem and is among the most efficient energy systems for delivering thermal energy [3]. DH efficiently generates heat centrally and distributes it through a network of insulated pipes, making it a scalable and environmentally sound alternative to conventional heating methods [4]. The International Energy Agency (IEA) expects DH systems to supply heat to 350 million buildings by 2030 [5], meeting 20% of global space heating needs; their role in sustainable urban development is undeniable.

However, fully realising the potential of DH systems requires overcoming several operational challenges, one of which is the frequent occurrence of operational faults in DH substations. These faults not only affect the individual substations but can also disrupt the performance of the entire network. With the shift towards Fourth-Generation District Heating (4GDH) systems [6], emphasising lower operating temperatures, the margins for error decreases [7]. Studies show that between 43% and 75% of the studied DH substations operate sub-optimally due to faults [8,9], and operators in major cities, such as Copenhagen, report that over 50% of their substations run sub-optimal due to faults [10]. The impact of faulty substations is significant; for instance, a large faulty substation can cost up to 100,000 euros per year of additional network pumping power [11]. The lack of automated Fault Detection and Diagnosis (FDD) methods aggravates this issue, enabling faults to remain undetected and unresolved for months or years.

FDD strategies, including data-driven approaches [12], predictive maintenance techniques [13], and deep-learning methods such as probabilistic load forecasting in integrated energy systems [14], provide a cost-effective and less disruptive alternative to traditional annual maintenance audits [9,15]. Complementing these strategies, the growing deployment of automatic heat meters in DH networks presents an opportunity to develop and implement data-driven solutions. By collecting and analysing operational data, these meters could help reduce DH customers' energy consumption by an average of 14% [16]. Nevertheless, some significant obstacles remain: the lack of labelled data [17] and current data collection efforts are primarily focused on billing purposes rather than FDD analysis. For instance, DH datasets typically include primary-side hourly measurements, such as supply temperature, return temperature, flow rate, and derived metrics like energy consumption. However, high-frequency or secondary-side information, such as temperatures, flow rates, and indoor settings, or additional data, including substation layouts and geographical locations, are often non-existent, not included or confidential, restricting the potential for in-depth or spatial analysis. As a result, there is a growing need for tailored methods specifically designed for the intricacies and data constraints in DH systems.

In this paper, we introduce HEAT (*Hierarchical-constrained Encoder-Assisted Time series clustering*), a novel approach for fault detection in DH. HEAT consists of two main phases: (1) *Network Topology Approximation*, which estimates the DH network structure to enable spatial analysis while preserving sensitive information, and (2) *Local Fault Detection*, which identifies abnormal behaviour of neighbouring DH substations using modified z-scores based on the Median Absolute Deviation (MAD), eliminating the need for labelled data. Our method leverages a constraint-based hierarchical clustering technique that integrates domain knowledge through a constraint matrix and can enforce minimum and maximum cluster size constraints using soft penalties, preventing the formation of excessively large clusters while encouraging the merging of smaller ones. This structured clustering approach enhances the identification of local performance anomalies, improves fault detection, and preserves data privacy by avoiding exposure to sensitive information such as geographic locations or detailed network layouts.

The structure of the paper is as follows: Section 2 reviews related work, Section 3 describes data acquisition, Section 4 details the proposed methodology, Section 5 presents the results and discussion, and Section 6 concludes the study.

2. Related work

Fault detection in DH systems heavily relies on global threshold-based methods or manual inspections [12,18]. Global comparison methods analyse all substations collectively to detect deviating substations [8,19,20], and individual methods, such as regression, set fixed thresholds for each substation [21,22]. While there are benefits of using such methods, they also introduce limitations. For instance, global approaches may highlight the extreme cases; they overlook the impact of local conditions, such as variations in supply temperature across different regions of the network, and therefore miss a large share of the subtle faulty substations. Moreover, Regression-based methods can fail when the model is built on data already affected by a fault, causing the faulty state to be normalised and ultimately go undetected.

Clustering methods have emerged as a powerful tool for analysing DH system behaviour, offering an alternative to traditional anomaly detection. Algorithms such as k-means (kM) and k-shape (kS) have been widely applied to uncover consumption patterns [23–25], detect outliers [26,27], classify substations based on heat load characteristics [28], and previous work has demonstrated the effectiveness of localised analysis [29]. However, a key challenge arises when conventional clustering, which minimises the total within-cluster distance, at higher values of k tends to isolate individual substations into singleton clusters rather than forming more balanced clusters that mirror the natural DH network structure. This cluster solution may compromise the interpretability and practical relevance, undermining the potential for meaningful intra-cluster comparisons. For instance, singleton clusters undermine intra-cluster analysis as there is no local neighbourhood to compare, while larger clusters can become too heterogeneous, widening intra-cluster spread and reducing sensitivity to subtle anomalies. To address these issues, a more structured clustering approach is required, one that imposes constraints to achieve balanced groupings while preserving the natural operational relationships among substations.

Constrained clustering has gained attention since the early 2000s for its ability to incorporate domain-specific knowledge to improve clustering outcomes [30,31]. Early work by Wagstaff et al. [32] introduced must-link and cannot-link constraints, laying the groundwork for constraint-based methods. Bar-Hillel et al. [33] further advanced the field with Relevant Component Analysis (RCA), offering a computationally efficient alternative to iterative methods such as those proposed by Xing et al. [34]. Similarly, Basu et al. [35] developed a semi-supervised framework that leverages a few pairwise constraints, along with an active learning strategy for selecting the most informative ones, leading to improved accuracy in high-dimensional settings. Recent advancements have focused on incorporating constraints within Hierarchical Agglomerative Clustering (HAC). Despite its higher computational cost compared to methods like kM, HAC offers the unique advantage of generating a dendrogram. This tree-like diagram visually represents the nested structure of the data, thereby enhancing the clarity and interpretability of the clustering process, a feature essential for crucial energy systems like DH. Additionally, HAC is well-suited to certain engineering contexts, such as DH networks, where substations initially function as isolated units that can later be grouped based on neighbours, streets, zones, and so on. However, when introducing constraints into HAC, it is crucial not to cut the dendrogram prematurely, as doing so may disrupt the natural hierarchical relationships within the data. For example, Davidson and Ravi [36,37] extended constrained clustering to HAC by integrating must-link and cannot-link constraints and addressed computational challenges through the introduction of a γ constraint. Chong and Lee [38] introduced a flexible framework that combines hard and soft constraints, using fuzzy AHP to manage uncertainty in semi-supervised clustering. In another development, Ambroise et al. [39] proposed an adjacency-constrained HAC tailored for genomic data, achieving quasi-linear complexity by limiting the clustering scope to adjacent objects. Most recently, Mauduit and Simonetto [40] employed a graph coarsening approach with optimal cuts to incorporate structural constraints in HAC, significantly enhancing constraint satisfaction and clustering cost in semi-supervised tasks.

Table 1
Summary of related work categories and representative studies.

Category	Representative studies
DH FDD review	van Dreven et al. [12], Rafati and Shaker [13] and Neumayer et al. [17]
DH threshold-based	Gadd and Werner [18], Månsson et al. [19], Månsson et al. [8] and Jangsten et al. [20]
DH regression methods	Theusch et al. [21] and Calikus et al. [22]
DH clustering-based approaches	Gianniou et al. [23], Tureczek et al. [24], Hong & Yoon [25], Xue and Zhou [26], Koussa et al. [27], Calikus et al. [28] and van Dreven et al. [29]
Classical constrained clustering	Bradley et al. [30], Wagstaff et al. [32], Bar-Hillel et al. [33], Xing et al. [34] and Basu et al. [35]
Constrained hierarchical clustering	Davidson and Ravi [36,37], Chong and Lee [38], Ambroise et al. [39] and Mauduit and Simonetto [40]

While prior research has predominantly employed must-link and cannot-link constraints, our HEAT method offers a more flexible approach for time series data by introducing adaptive soft constraints that enable should-link and should-not-link relationships. This approach balances the flexibility of freely clustering with rigid must-link constraints, preventing the premature termination of the dendrogram (dead-end branching) and altering the clustering process by encouraging or discouraging merges. By utilising adaptive penalties, HEAT dynamically regulates minimum and maximum cluster sizes, resulting in more uniformly sized clusters. Additionally, our method employs a constrained matrix for point-to-point penalties, thereby providing a more nuanced solution towards incorporating domain knowledge into the clustering process. Finally, in the second stage of HEAT, we use the clustering outcome for a more contextual intra-cluster anomaly detection of DH substations using MAD and modified z-scores. A consolidated comparison of related work is provided in Table 1.

3. Data acquisition

The proposed approach was validated using operational data collected from a DH network located in southern Shandong Province, China. Due to the sensitive nature of the data, it remains confidential and is not available to the public. The dataset consists of time series sensor data from 248 substations, logged every 5 min during the heating season of January 2024. As a result, each substation provides 8928 time-stamped samples (31 days × 288 samples per day), capturing the network's performance under outdoor temperatures ranging from −9 °C to 11 °C. The dataset includes common primary measurements, such as supply- and return temperatures (in °C), flow rate (in L/min), and calculated values such as energy consumption (kWh), water volume (m^3) and temperature differences (ΔT). The primary supply-temperature profiles in the DH network are centrally controlled (network μ = 86.23 °C, σ^2 = 80.23 $°C^2$). HEAT uses these supply-temperature signals directly, projected into a lower-dimensional latent space, to approximate network topology, while the ΔT series (network μ = 43.84 °C and σ^2 = 74.83 $°C^2$) is reserved for our intra-cluster anomaly analysis. The central control of supply temperatures and short analysis period ensure the data are relatively stationary with limited diurnal or seasonal components. For a detailed list of sensor measurements, see Table 2. Furthermore, the dataset includes partial labelling,

Table 2
Primary side time series data characteristics of real-world DH substations.

Feature	Abbr.	Type	Description
Timestamp	TS	Datetime	Recorded time of measurement
Primary supply Temp.	T_s	Continuous	Temperature of primary supply line (°C)
Primary return Temp.	T_r	Continuous	Temperature of primary return line (°C)
Primary flow	Q	Continuous	Flow rate of primary circuit (L/min)
Outdoor Temp.	T_{out}	Continuous	Ambient temperature outside the system (°C)
Temp. difference	ΔT	Continuous	Difference between T_s and T_r (°C)
Energy consumption	E	Continuous	Total energy used by a substation (kWh)
Volume	V	Continuous	Total volume of water circulated (m^3)

including fault types such as High Heat Curve (HHC), Wrong Sensor Placement (WSP), Large Valve (LV), Large Secondary Leakage (LSL), and normal behaviour [8]. Fault annotations are provided by the DH utility and supplemented with insights through independent analysis by a domain expert.

4. Proposed method

Our approach, HEAT, aims to identify two distinct types of anomalies within a DH network, supply temperature and performance anomalies, through two phases. In the first phase of our methodology, in Section 4.1, we construct a relative network topology that uses the supply temperature profiles as a distance metric to approximate the spatial and operational relationships between substations. Anomalous supply temperature substations are isolated into singleton clusters, while we group substations with similar temperature profiles together as they suggest neighbouring substations. Building upon this network topology, in Section 4.2, we implement an anomaly detection strategy that focuses on the substation performance. Specifically, we assess each substation's operational behaviour in relation to its immediate neighbours.

4.1. Constrained clustering

We use the primary supply temperature time series, specifically the heat losses over time, as a distance metric to approximate the relative network topology. The primary supply temperatures reaching consumer substations are influenced by several factors contributing to heat loss, such as flow rate, pipe length and diameter, and initial water supply temperature [41]. Since a substation's location in the network dictates a unique combination of these factors, as closely neighboured substations tend to share similar characteristics, we can approximate each substation's relative position to the heat source and its neighbouring substations. For further details about using supply temperatures for DH topology approximation, we refer the readers to [29].

However, time series data, such as supply temperatures, are inherently high-dimensional, with each time point contributing to the dataset's complexity. This high dimensionality can obscure underlying patterns, as distances between points in high-dimensional space tend to concentrate, degrading the performance of clustering algorithms [42]. In addition, raw time series data include noise and redundant information that can lead to spurious clusters or obscure meaningful groupings [43,44]. Dimensionality reduction techniques aim to extract the most informative features by compressing the data into a lower-dimensional latent space [45]. This compression helps mitigate noise and redundancy and enhances the interpretability and efficiency of subsequent clustering methods. By transforming the data into a more compact form, we can ensure that the clustering algorithm focuses on the essential patterns and structures in the time series.

To encode the time series data into a lower-dimensional latent space representation, we use a convolutional autoencoder (CAE). Autoencoders (AEs) are neural networks used for unsupervised learning, aiming to compress input data into a latent space and reconstruct it. In our implementation, we adopt convolutional layers instead of fully connected ones, as past research has shown that they are well-suited for capturing local temporal dependencies in time series data [46–48]. Max-pooling layers reduce the temporal resolution, allowing the network to learn hierarchical features at different temporal scales, which is beneficial for time series analysis.

Fig. 1 demonstrates the CAE architecture we used, which is tailored to our DH use case. The CAE architecture is designed to compress and reconstruct one-dimensional DH time series data. The encoder comprises four convolutional layers with progressively decreasing filter sizes (76, 64, 32, and 16). Each convolutional layer is followed by batch normalisation to stabilise and accelerate training, a LeakyReLU activation function ($\alpha = 0.1$) to model non-linear dependencies, dropout (rate 0.3) for regularisation, and max pooling to downsample the temporal resolution. The compressed representation is projected into an eight-dimensional latent space via a convolutional layer. Under the manifold hypothesis, undercomplete AEs typically enforce a bottleneck size that is a small fraction of the input dimension to capture intrinsic structure while suppressing noise [49–51]. Systematic approaches can also estimate a minimum viable latent dimension [52]. Although deriving an optimal d is beyond this study's scope, our choice of latent size follows these guidelines. The decoder reverses the encoder structure, employing transposed convolutional layers with symmetric filter configurations to restore the temporal resolution and reconstruct the original time series.

Before discussing the proposed clustering constraints mechanism, it is important to note that unbalanced clustering can lead to impractical solutions. In a previous study [29], the authors observed that when the number of clusters increased, the algorithm favours separating individual samples into individual clusters. While identifying true outliers is important, creating clusters of only one sample may not provide meaningful insights for further analysis. Enforcing balanced clusters helps prevent such trivial assignments and encourages forming groups that better represent the underlying structure in the DH network. Moreover, balanced cluster distributions are crucial not only to avoid unnecessary singleton clusters but also to prevent excessively large clusters from diluting statistical measures. For instance, many samples can smooth out the variability needed for modified z-score outlier detection, potentially obscuring true outliers within that cluster. By enforcing more balanced cluster sizes, we ensure that the statistical characteristics remain sensitive enough to detect the faulty substations more robustly.

Therefore, we propose an adaptive penalty mechanism for the HAC linkage function that modifies the inter-cluster distance based on minimum cluster size $C_{\min}$, maximum cluster size constraints $C_{\max}$, or a provided constraint matrix M. This mechanism ensures that cluster sizes remain within a desirable range or follow specific point-to-point constraints as required. The minimum and maximum size constraints are applied at the cluster-to-cluster level, where penalties are adaptively enforced based on how much the cluster sizes deviate from $C_{\min}$ or $C_{\max}$. Conversely, the constraint matrix M penalises or encourages merges based on individual substation relationships at a point-to-point level, with the total penalty for merging two clusters being the sum of all pairwise constraints between their respective members. Note that a penalty is negative to favour a merge, while positive to discourage a merge. Prior to clustering, we normalise the network's pairwise distance matrix D by its maximum entry $D_{\text{norm}} = D \ / \ \max_{i,j} D_{ij}$.

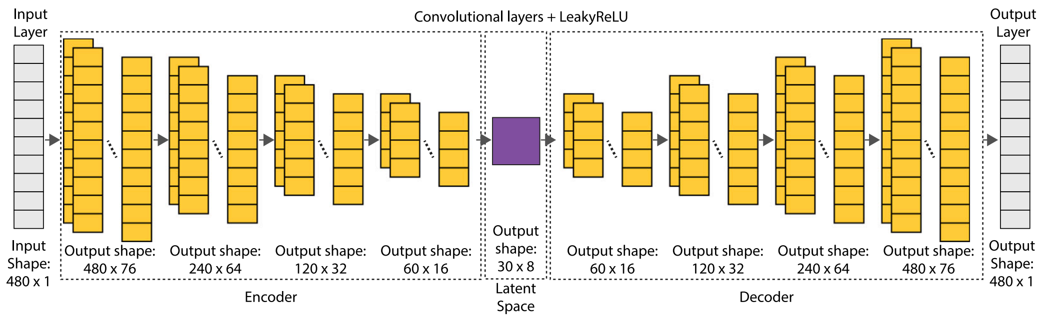


Fig. 1. The CAE model architecture used in this study.

This ensures all distances lie in $[0,1]$, so penalty parameters, such as $w_{\min}, w_{\max} \in [0,1]$ retain the same semantic interpretation (fractions of the network diameter) across different DH systems, improving transferability of constraints or penalties.

Formally, for $C_{\min}$, we adjust the distance d_{ij} between two clusters I and J as follows:

$$d_{ij} \leftarrow d_{ij} + w_{\min} \times \left(C_{\min} - (|I| + |J|)\right), \quad \text{if } |I| + |J| < C_{\min}, \tag{1}$$

where $|.|$ denotes cardinality and $w_{\min}$ the weight that controls the strength of the penalty. This ensures that clusters smaller than $C_{\min}$ become more favourable to merge unless they reach the adequate size.

Similarly, if the combined size of clusters I and J exceeds the maximum cluster size $C_{\max}$, we impose an adaptive maximum penalty proportional to the excess size as follows:

$$d_{ij} \leftarrow d_{ij} + w_{\max} \times \left(|I| + |J| - C_{\max}\right), \quad \text{if } |I| + |J| > C_{\max}, \tag{2}$$

where $w_{\max}$ controls the strength of the penalty, making merges between overly large clusters less favourable.

In addition to the size-based constraints, domain-specific knowledge is integrated through a constraint matrix M. Each element $M[i,j]$ represents a penalty or reward for merging specific points. If two points i and j are rewarded for merging, the penalty is negative. In contrast, the penalty is positive if merging is penalised. The distance is updated as follows:

$$d_{ij} \leftarrow d_{ij} + \sum_{\substack{p \in \text{cluster}_i \\ q \in \text{cluster}_j}} M[p,q] \tag{3}$$

The matrix M allows for the soft enforcement of must-link and cannot-link constraints between individual points, guiding the clustering process based on domain-specific relationships between substations. These constraints are applied softly, influencing the clustering process but not strictly enforcing merges or splits. However, suppose strict must-link and cannot-link constraints are desired; assigning a sufficiently high penalty value and selecting an appropriate cut-off point in the dendrogram ensures adherence to these constraints. Algorithm 1 outlines the linkage function that integrates the soft constraints into the hierarchical clustering process, thereby enabling more flexible clustering while guiding the formation of balanced clusters.

HEAT seeks to optimise two primary metrics: the mean Dynamic Time Warping (DTW) distance (μ_{dtw}) and the mean variance in cluster sizes (σ^2_{size}). The μ_{dtw} within clusters measures cluster compactness (intra-cluster similarity), capturing the temporal similarity between time series. A lower μ_{dtw} indicates that the time series in a cluster are more temporally similar, reflecting a higher degree of cohesion within the cluster. While lower σ^2_{size} ensures that clusters are more evenly distributed, avoiding scenarios where some clusters are disproportionately large or small.

We combine these two objectives using the geometric mean (G_p), formally defined as:

$$G_p = \sqrt{\mu_{dtw} \times \sigma^2_{size}}, \tag{4}$$

Algorithm 1 HEAT Linkage Function

Require: Distance matrix D, Min cluster size $C_{\min}$, Max cluster size $C_{\max}$, Min weight w_{min}, Max weight w_{max}, Constraint matrix M
Ensure: Linkage matrix L

Initialize clusters: $C \leftarrow \{\{i\} \mid i = 0, \dots, n-1\}$, where n is the number of samples
Initialize sizes: $S[i] \leftarrow 1$ for all $i \in \{0, \dots, n-1\}$
Set linkage matrix $L = \{\}$
while $|C| > 1$ **do**
  Set $d_{\min} \leftarrow \infty$
  Initialize optimal clusters pair: $I^* \leftarrow \texttt{None}, \quad J^* \leftarrow \texttt{None}$
  for all pairs of clusters (I, J) **do**
    $d_{ij} \leftarrow dist(I, J)$
    if $M \neq \texttt{None}$ **then** ▷ Pairwise constraints
      Update: $d_{ij} \leftarrow \max\left(d_{ij} + \sum_{p \in I} \sum_{q \in J} M[p,q], 0\right)$
    end if
    if $C_{\min} \neq \texttt{None}$ and $(|I| + |J|) < C_{\min}$ **then** ▷ Minimum cluster size constraint
      Update: $d_{ij} \leftarrow \max\left(d_{ij} + w_{min} \times (C_{\min} - (|I| + |J|)), 0\right)$
    end if
    if $C_{\max} \neq \texttt{None}$ and $(|I| + |J|) > C_{\max}$ **then** ▷ Maximum cluster size constraint
      Update: $d_{ij} \leftarrow d_{ij} + w_{max} \times ((|I| + |J|) - C_{\max})$
    end if
    if $d_{ij} < d_{\min}$ **then**
      Set $d_{\min} \leftarrow d_{ij}$ and update $(I^*, J^*) \leftarrow (I, J)$
    end if
  end for
  Merge: $k \leftarrow I^* \cup J^*$
  Update sizes: $S[k] \leftarrow S[I^*] + S[J^*]$
  Append $(I^*, J^*, d_{\min}, S[k])$ to L
  Update distances: $D(k, m) = \text{dist}(k, m), \quad \forall m \neq k$
  Remove: $C \leftarrow C \setminus \{I^*, J^*\}$
  Add new: $C \leftarrow C \cup \{k\}$
end while
return Linkage matrix L

The G_p is used as it ensures that μ_{dtw} and σ^2_{size} are optimised simultaneously. A low value in either component will result in a lower overall score, preventing the dominance of one metric over the other. Thus, this formulation allows us to achieve both compact clusters (low μ_{dtw}) and balanced cluster sizes (low σ^2_{size}).

4.2. Fault detection

We calculate MAD z-scores to robustly measure dispersion (intra-cluster), with modified z-scores [53] similar to [29]. The local neighbouring substations with mean z-scores below the threshold (single-sided) of two standard deviations (2σ) from the intra-cluster distribution mean, which indicates average neighbourhood performance, are flagged as faulty.

Specifically, for each cluster C_k, let

$$z_i = \frac{0.6745\left(x_i - \text{median}\{x_j : j \in C_k\}\right)}{\text{MAD}\{x_j : j \in C_k\}}$$

be the MAD z-score of substation i, where x_i is the measured ΔT (Eq. (5)). We then compute, within each cluster, the sample mean μ_k and standard deviation σ_k of $\{ z_i : i \in C_k \}$. Any node i with $z_i < \mu_k - 2\sigma_k$

is flagged as faulty. In other words, the threshold for cluster k is set at approximately its own 2σ lower tail, making it directly adaptive to that cluster's spread and location. Compared to a fixed, global z-score cutoff, this adjustment aims to improve sensitivity to moderate deviations in clusters of differing variance.

To assess substation performance, we focus on the primary ΔT, a key performance indicator that reflects heat transfer efficiency within a substation. Moreover, substation temperature readings, such as return temperature and ΔT, serve as reliable indicators for detecting faults [54,55]. ΔT is formally defined as:

$$\Delta T = T_s - T_r, \tag{5}$$

where T_s is the primary supply temperature and T_r the primary return temperature, with higher ΔT values indicating better heat extraction, i.e., better performance. However, since the initial supply temperature influences ΔT, its absolute value can vary under different conditions, limiting the effectiveness of a goal analysis and necessitating a more contextual approach.

4.3. Experimental setup and evaluation metrics

We conducted several experiments to validate our approach stages, including clustering and fault detection. In the first experiment, we benchmarked our HEAT algorithm against several industry-standard clustering techniques [12]. All experiments were conducted by a single researcher on a standard desktop workstation (Intel Core i7-9700 CPU, 16 GB RAM, without GPU acceleration). Most of the computing time is consumed by the CAE (100 epochs). In contrast, the HAC stage scales approximately as $\mathcal{O}(n^3)$ but was computationally negligible in practice for the studied network of 248 substations and the anomaly detection step was near-instantaneous.

We use cluster implementations available in SciPy [56], tslearn [57], and TensorFlow [58], including:

- HAC [59] and CAE-HAC,
- Time Series kM (TS kM) [60,61] and CAE-kM,
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [62], including its variants Hierarchical DBSCAN (HDBSCAN) [63] and DENsity-based CLUstEring (DENCLUE 2.0) [64],
- Spectral Clustering (SC) [65] and kS [66].

Their performance is measured using the metric G_p. Traditional constrained clustering methods rely on must-link and cannot-link constraints to maintain pairwise relationships rather than explicitly managing cluster sizes. Our work prioritises balanced cluster sizes while maintaining interpretable results, therefore, we excluded these methods from our comparisons.

To robustly evaluate the constraint matrix, we conducted multiple trials where arbitrary must-link and cannot-link constraints were assigned to substation pairs. In each trial, we randomly assign 10% must-link and 10% as cannot-link constraints. We chose this proportion to provide sufficient guidance to influence the clustering outcome in a controlled way while avoiding an overconstrained setup. We measure the constraint performance using the Constraint Satisfaction Rate (CSR). A must-link constraint is satisfied if the two assigned data points (substations) belong to the same cluster, while a cannot-link constraint is satisfied if the two points are placed in different clusters. The CSR is defined as:

$$CSR = \frac{C_{satisfied}}{C_{total}}, \tag{6}$$

where $C_{satisfied}$ is the number of must-link and cannot-link constraints the clustering correctly follows, and C_{total} is the total number of all constraints applied in the experiment. The CSR value ranges from 0 to 1, where 1 indicates all constraints are adhered to, while 0 means none of the constraints were followed.

We report the mean CSR along with the 95% confidence interval (CI) to assess the variability and reliability of the constraint enforcement. We use a baseline test of arbitrary assigning samples to clusters. If our experiments show a significantly higher CSR compared to the baseline, it suggests that the constraints are influencing the clustering process.

In the second experiment, we evaluated fault detection by comparing HEAT with several baseline network analysis strategies:

- **Global Analysis:** All substations were assessed concurrently in the fault detection stage (as one single cluster) to establish an overall performance benchmark.
- **Localised Group Analysis:** The DH network was partitioned into arbitrarily and equally distributed groups to investigate the impact of localised clustering on fault detection sensitivity.
- **Comparative Analysis:** HEAT was directly compared against the second-best clustering algorithm from our clustering analysis, highlighting its improvements in forming meaningful clusters.

We evaluate the outcome using statistical tests to rigorously assess the systematic effects present in our data and quantify the magnitude of the overall effect size. Specifically, we applied the two-sample t-test assuming unequal variances (Welch's t-test [67]) to verify that observed differences reflect systematic effects rather than random fluctuations. The resulting p–values ($p < \alpha$, where $\alpha = 0.05$) indicate statistically significant differences, confirming that these patterns are highly unlikely to arise by chance. We also calculated Cohen's d to quantify the effect size, reflecting the practical differences [68]. This approach allows us to evaluate the trade-offs between network granularity and overall system robustness, determine whether the observed effects are statistically significant and understand the practical implications by measuring their strength. Moreover, these experiments provide a holistic evaluation of HEAT's fault detection performance.

Finally, fault detection performance is measured using sensitivity and specificity. Sensitivity, or the true positive rate, reflects the model's ability to identify actual anomalies correctly and is defined as:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

where TP and FN denote true positives and false negatives, respectively. TP are faulty substations correctly identified, and FN are faulty substations missed. Specificity, or the true negative rate, quantifies the ability to identify negatives correctly and is expressed as:

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{8}$$

where TN and FP represent true negatives and false positives, respectively. TN are correctly unflagged normal substations, and FP is normal substations incorrectly flagged.

To provide a balanced evaluation of the model's performance, we also calculate the geometric mean (G_d), which combines sensitivity and specificity into a single metric:

$$G_d = \sqrt{\text{sensitivity} \times \text{specificity}} \tag{9}$$

The G_d metric is beneficial because it penalises models that perform well on one metric but poorly on the other. A high G_d value can only be achieved when both sensitivity and specificity are high, indicating that the model performs well across both positive and negative classes.

5. Results and discussion

In the following section, we present and discuss our results. Since our method compromises a two-step phase, we initially present the clustering results in Section 5.1. In Section 5.2, we present the fault detection results.

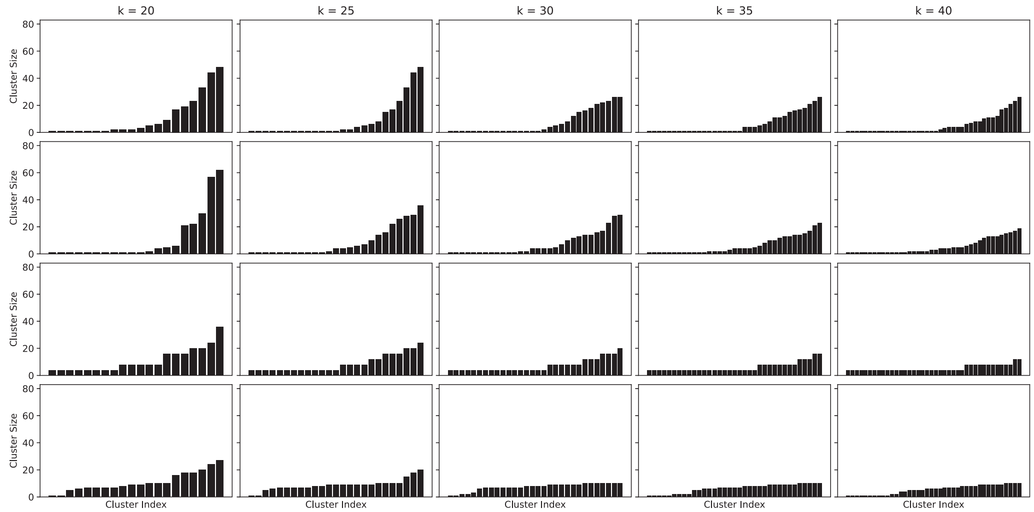


Fig. 2. Cluster size distributions for regular HAC, CAE-HAC, HEAT (min cluster size 6), and HEAT (max cluster size 10) at different values of k.

5.1. Clustering

We evaluated the performance of HEAT on time series data from DH substations. The HEAT method uses a CAE for dimensionality reduction and integrates soft constraints to guide the clustering process of HAC. We conducted two experiments to assess the HEAT. In the first experiment, we compared HEAT with regular HAC and CAE-HAC. In the second experiment, we compare HEAT against industry standard time series clustering methods, such as TS k-M, CAE-TS k-M, DBSCAN (HDBSCAN and DENCLUE 2.0), SC, and kS, to highlight its overall effectiveness. Our aim of the clustering process was to have tightly packed and coherent clusters and more equally balanced cluster sizes.

5.1.1. Comparison of HEAT with hierarchical clustering variants

Our initial experiment compares HEAT against the standard HAC, which uses the original high-dimensional time series data, and a CAE-HAC implementation, using the lower-dimensional representations obtained from the CAE. HEAT combines the CAE with integrating soft constraints in the linkage function of HAC. This experimental design allows us to examine the effects of HEAT and the CAE on hierarchical clustering.

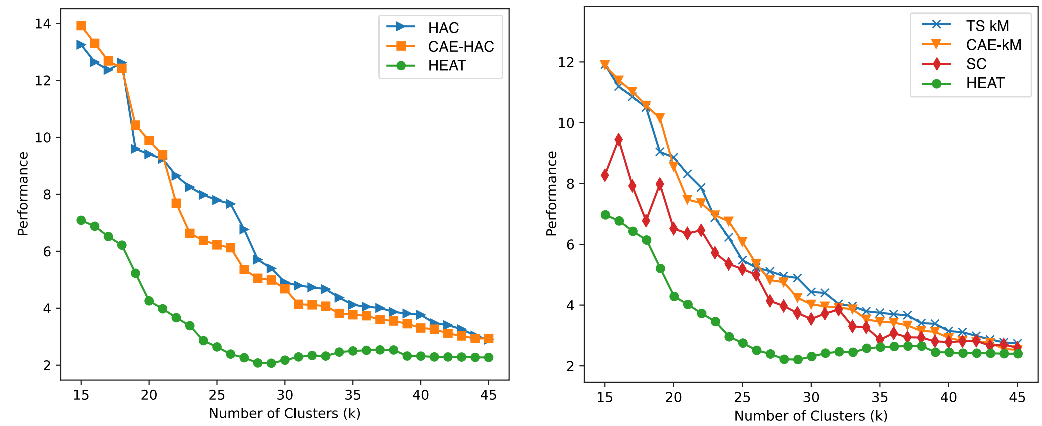
In Fig. 2, we display the cluster size distributions for different numbers of k across the methods. Specifically, we set the minimum cluster size to 6 and the maximum cluster size to 10. These values were derived from domain-specific considerations. For instance, a minimum of 6 substations is necessary to ensure reliable comparisons. In comparison, a maximum of 10 (representing neighbouring substations in a street) prevents overly large clusters, thereby maintaining a balance between granularity and robustness in analysis. At $k = 20$, while some violations of the size constraints occur, HEAT still produces clusters that are noticeably more balanced in terms of both minimum and maximum size. As k increases, HEAT yields clusters that adhere more closely to the imposed bounds. Notably, a sufficiently large k is required to allow the constraints to be effectively enforced; if k is too small, it is impossible to fully adhere to the imposed minimum and maximum cluster size bounds. A higher number of clusters k facilitates effectively enforcing the constraints. In contrast, even at elevated k values, both regular HAC and CAE-HAC exhibit significant size disparities, which is as intended as they aim to minimise the total within-cluster distance and not enforce uniformity in cluster sizes. We can observe in Fig. 2 that several clusters are becoming overly large in HAC and CAE-HAC for the various k clusters, a limitation in [29].

From Fig. 3(a), we observe that HEAT outperforms both HAC and CAE-HAC in terms of cluster performance G_p (defined in Section 4.1). Our results reveal that CAE-HAC slightly outperforms regular HAC. This observation indicates that the low-dimensional representations produced by the convolutional AE capture essential features that enhance cluster cohesion and separation. The modest improvement of CAE-HAC over HAC suggests that incorporating an AE can be beneficial for clustering high-dimensional time series data in DH systems by reducing noise and emphasising relevant patterns. However, a significant performance increase is observed for HEAT. HEAT minimises intra-cluster distances, thereby improving cluster cohesion, and maintains more uniform cluster sizes. This indicates that introducing soft constraints does not compromise the local cohesion of clusters; rather, it strikes an effective balance between preserving intrinsic data structure and enforcing desirable cluster size properties. For instance, at $k = 28$, HEAT reaches its best G_p of approximately 2.1, a reduction of approximately 58.4% relative to CAE-HAC. Since lower G_p indicate better performance, it demonstrates a substantial improvement. Even at higher k ($k = 39$), HEAT still showed better performance with a substantial reduction of approximately 34.4% relative to CAE-HAC. It is important to note that as k increases, performance typically improves because the clustering objective, measuring both intra-cluster cohesion and cluster size dispersion, becomes more straightforward to optimise. In the limit, where each substation forms its cluster, the distance and size distribution component is fully optimised; however, the balance between cohesion and size uniformity remains critical due to the effects of the G_p, and HEAT effectively navigates this trade-off.

5.1.2. Comparison of HEAT with conventional methods

We further compared HEAT with several industry standard clustering algorithms: TS kM, which is directly applied to the raw time-series data; CAE-TS kM, which utilises the lower-dimensional representations obtained from the CAE; and SC, which leverages the eigenvalues of the similarity matrix derived from the original time series data. Preliminary experiments with additional density-based methods (e.g., DBSCAN, HDBSCAN, and DENCLUE 2.0) and a centroid-based algorithm kS revealed suboptimal performance. In particular, at higher k values, these methods often fail to identify a sufficient number of clusters, resulting in a proliferation of empty clusters. Consequently, we excluded them from further analysis.

Fig. 4 presents the cluster size distributions for the evaluated methods across different values of k. With the slight exception of SC, we observe a similar trend: HEAT can produce the best balance in cluster size distribution, while the other methods yield a high imbalance; many standard methods tend to produce either overly large clusters or many singletons as k increases, since the clustering objective can often be minimised by isolating individual samples into singleton clusters rather than partitioning a large cluster into more balanced sub-clusters.



(a) G_p of cluster cohesion and size uniformity for hierarchical clustering variants.

(b) G_p of cluster cohesion and size uniformity for industry-standard methods.

Fig. 3. Performance results expressed as the G_p of cluster cohesion and size uniformity: (a) compares hierarchical clustering variants, (b) contrasts HEAT with industry-standard methods.

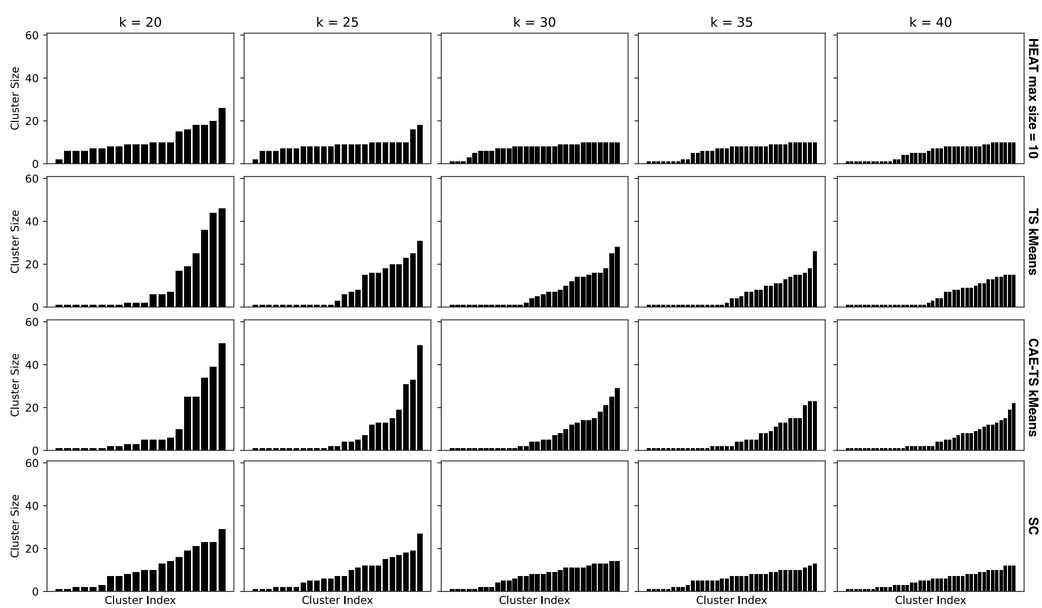


Fig. 4. Cluster size distributions for HEAT (max cluster size 10), TS k-means, CAE-TS k-means, and SC at different values of k.

From Fig. 3(b), we can observe that HEAT is consistently outperforming all compared methods across every k value. However, while the other methods generally outperform regular HAC, indicating that HAC has inherent limitations concerning our objective, our proposed method, HEAT, compensates for these shortcomings. Interestingly, TS k-M and CAE-k-M exhibit very similar performance, suggesting that the inclusion of the CAE yields only a modest improvement. Notably, although SC achieved relatively balanced cluster sizes, as shown in Fig. 4, this balance was attained at the expense of cluster compactness. The trade-off underscores the inherent challenge in clustering: optimising one criterion, such as size uniformity, may compromise another, like intra-cluster cohesion. Nevertheless, SC was second-best and came close to the performance of HEAT at some k (e.g., $k = 35$). However, SC does show the most variability across k, especially at the start, suggesting it struggled to find a proper balance when k clusters is low. As k increases, SC becomes better at optimising the overall clustering objective.

The results demonstrate that HEAT effectively addresses the challenges of clustering high-dimensional time-series data by leveraging the CAE and soft constraints. The CAE captures essential patterns in the data, reducing dimensionality and facilitating the clustering process. Concurrently, soft constraints are generally more effective and appropriate than hard rules, especially in scenarios where cluster size limits are enforced. While hard rules rigidly enforce predefined boundaries, they can do so at the expense of disregarding more meaningful relationships within the data. Another key advantage of HEAT is its flexibility to guide the clustering process. Although constrained clustering techniques exist, often implemented with methods such as kM [30], HEAT leverages hierarchical clustering, which offers distinct benefits. Admittedly, HAC is computationally more intensive than some partitioning methods; however, this trade-off is less critical when the number of samples (e.g., DH substations) is limited. For instance, HAC produces deterministic results across multiple runs, ensuring the reproducibility of the clustering structure, which is a crucial factor for critical energy systems such as DH networks. Moreover, the dendrogram produced by HAC visually represents the merging process, which enhances explainability by clarifying how clusters are formed and how data points relate to one another. The tree structure also facilitates multi-resolution analysis, enabling global and local insights into the data structure. For example, the method can be applied for fault detection or optimising supply temperatures across the network,

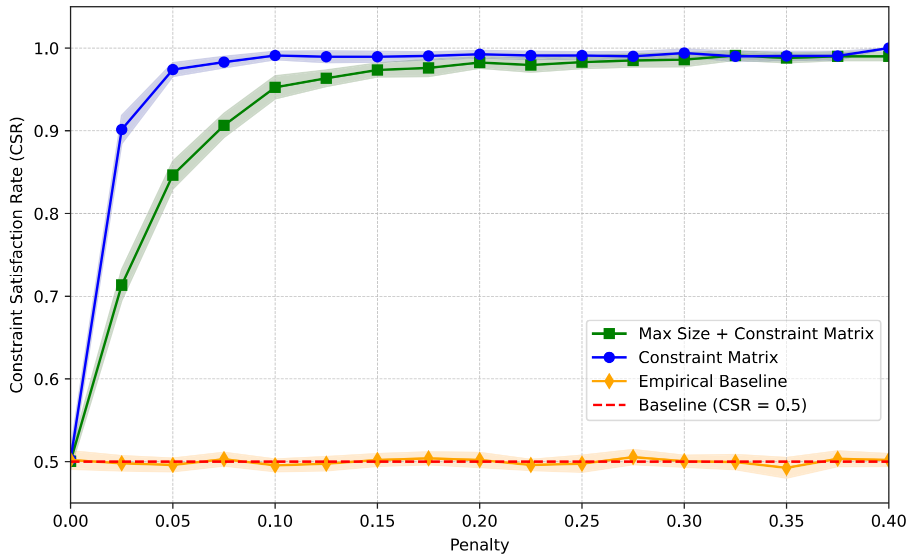


Fig. 5. The effect of penalty strength on the CSR for different must-link and cannot-link constraint settings with HEAT clustering. The solid lines represent the mean CSR across multiple experiments, while the shaded regions indicate the 95% CI. The red dashed line at CSR = 0.5 represents the theoretical baseline, followed by random clustering. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

delivering energy at greater efficiency. Additionally, the interpretability of the hierarchical structure and repeatability allow for the seamless integration of domain knowledge and expert insights, further enhancing the method's practical relevance. Finally, HEAT's integration of soft constraints introduces parameter flexibility, allowing a tunable balance between cluster cohesion and size uniformity to be adapted to specific application requirements. Together, these advantages render HEAT particularly well-suited for industrial use cases such as in DH systems.

5.1.3. Effects of the constraint matrix

The results in Fig. 5 demonstrate the effectiveness of the constraint matrix in guiding the clustering process. At the lower penalty introduced (0.025), the results are statistically significant with a large effect size ($T = 49.90$; $p < 0.001; d = 9.98$). The observed results can be explained by considering how the penalty term interacts with the clustering objective. In our proposed algorithm, the constraint matrix introduces an additional cost for violating the must-link and cannot-link relationships. At the lowest penalty (e.g., 0.0), the algorithm behaves similarly to an unconstrained clustering approach, so the CSR remains close to the baseline. However, as the penalty increases, any deviation from the imposed constraints incurs a greater cost, forcing the algorithm to adjust the cluster assignments to minimise these penalties. Specifically, with a small penalty (around 0.025 to 0.075), HEAT strongly prioritises constraint adherence over the natural, data-driven clustering structure. This leads to a rapid improvement in CSR as the algorithm reassigns data points to satisfy must-link and cannot-link relationships more consistently. As the penalty increases, the cost of violating any constraint becomes so high that the algorithm converges on a solution where nearly all constraints are satisfied, eventually reaching 1.0 for very high penalty values.

Introducing a maximum cluster size constraint adds another layer of complexity. At the lower penalty (0.025), the CSR reaches 0.7135, which is still significant with a large effect size ($T = 24.75$; $p < 0.001; d = 4.95$). We can observe in Fig. 5 that for lower penalty levels, the algorithm must balance between enforcing the relational constraints and respecting the cluster size limits, which can slightly delay the increase in CSR. For lower penalties (e.g., 0.05 to 0.15), the CSR suggests not all constraints are satisfied. Essentially, the maximum cluster size acts as an additional soft constraint that competes with the must-link and cannot-link requirements. Although this causes some variability (as reflected by the slightly higher CIs), the overall trend remains: increasing the penalty strength eventually forces the algorithm to comply with the constraints, compared to the unconstrained baseline case.

Overall, the results demonstrate a systematic effect of how the soft constraint matrix guides the clustering process by balancing the trade-offs between constraint adherence and intrinsic data structure. As the penalty increases, HEAT increasingly prioritises the satisfaction of must-link and cannot-link constraints over maintaining the unconstrained clustering structure, i.e., higher penalty values lead to greater costs for any constraint violations, thereby enforcing the constraints more strictly, eventually approximating or becoming hard constraints (Fig. 4 penalty = 0.40 results in a CSR = 1.0). While lower penalties provide gentle guidance, allowing the intrinsic data structure to determine the clustering outcome predominantly. This tunability is particularly advantageous in real-world applications where requirements may vary, allowing for a tailored balance that can respect the data's natural grouping and the valuable external information encoded in the constraints.

5.2. Fault detection

The fault detection phase builds on the clustering solution developed earlier. Given the scarcity of labelled data in DH systems, we assess a substation's performance relative to its peers (intra-cluster), where deviations may signal faults or suboptimal performance. Based on the performance results in Section 5.1, we continue our analysis with $k = 39$, providing a good balance by separating outlier substations and sufficient groups for further performance comparisons. To evaluate performance, we analysed the ΔT of each substation, as outlined in Section 4.2. Lower ΔT values indicate poorer performance. Our evaluation uses modified z-scores to quantify performance deviations and summarise the detection performance using sensitivity (the proportion of correctly identified faults) and specificity (the proportion of correctly identified normal substations). HEAT's performance was compared against several baseline methods: global detection (no clustering), second-best clustering SC, and detection using arbitrary equally distributed groups to measure the effects of smaller groups on detection rates. We summarise the results in Table 3.

Table 3
Comparison of fault detection methods including 95% CI.

Method	Sensitivity (%)	Specificity (%)	G_d (%)
Global	14.80 ± 1.99	89.91 ± 2.38	36.50 ± 4.00
SC	29.17 ± 1.72	94.55 ± 1.00	52.46 ± 1.29
Arbitrary equal clusters	58.30 ± 3.61	92.50 ± 1.67	73.40 ± 3.80
HEAT	**74.10 ± 1.99**	**95.50 ± 0.31**	**84.10 ± 1.80**

5.2.1. Performance of baselines

Global detection achieved a sensitivity of 14.80% ± 1.99 and specificity of 89.91% ± 2.38, respectively. These results indicate that global detection performs poorly in identifying faulty substations. While the method could identify a few faulty substations with the most pronounced deviations in ΔT, e.g., an incorrect valve, it struggled to detect more subtle faults, such as secondary leakages. In the case of the global comparisons, the aggregation of data results in an averaging effect, and substations with subtler impacting faults were masked by the relatively larger impacting faulty substations. While the specificity of the global analysis was relatively high, minimising unnecessary maintenance interventions, its poor sensitivity highlights its inadequacy for robust fault identification in DH systems, and many of the faults went unnoticed.

The SC baseline increased its sensitivity to 29.17% ($T = 6.78$; $p < 0.001$; $d = 4.8$) and specificity to 95.55% ($T = 2.96$; $p < 0.029$; $d = 1.8$) substantially with a large effect size as compared to the global baseline. Although forming more meaningful clusters appears to enhance fault detection performance, the method still suffers from a limitation as noted in [29] and seen in global analysis; the unequally balanced cluster sizes or overly large clusters constrain the overall effectiveness of the approach. Nevertheless, the substantial improvement in specificity suggests that the SC baseline effectively reduces the false positive rate, potentially decreasing unnecessary maintenance efforts instead of global analysis.

The arbitrary group's baseline, where substations are randomly assigned in equally sized groups, demonstrated a notable improvement compared to global analysis, with sensitivity increasing to 58.30% ± 3.61 and specificity to 92.50% ± 1.67, respectively. The effect of having smaller but equally distributed groups of substations for performance comparison, as opposed to global comparisons, is significant and showed a substantial increase in sensitivity ($T = 20.67$; $p < 0.001$; $d = 9.24$). Smaller groups likely mitigate the masking effect of pronounced faulty substations in global analysis. Moreover, it outperformed the second-best clustering algorithm SC in terms of sensitivity ($T = 9.01$; $p < 0.001$; $d = 6.37$), at the cost of a lower specificity (92.50%). Such improvements suggest smaller, balanced groups are helpful at capturing local anomalies that might be diluted in a cluster solution with size imbalances. However, the arbitrary nature of the groupings likely restricts the method's ability to achieve higher performance, particularly in specificity.

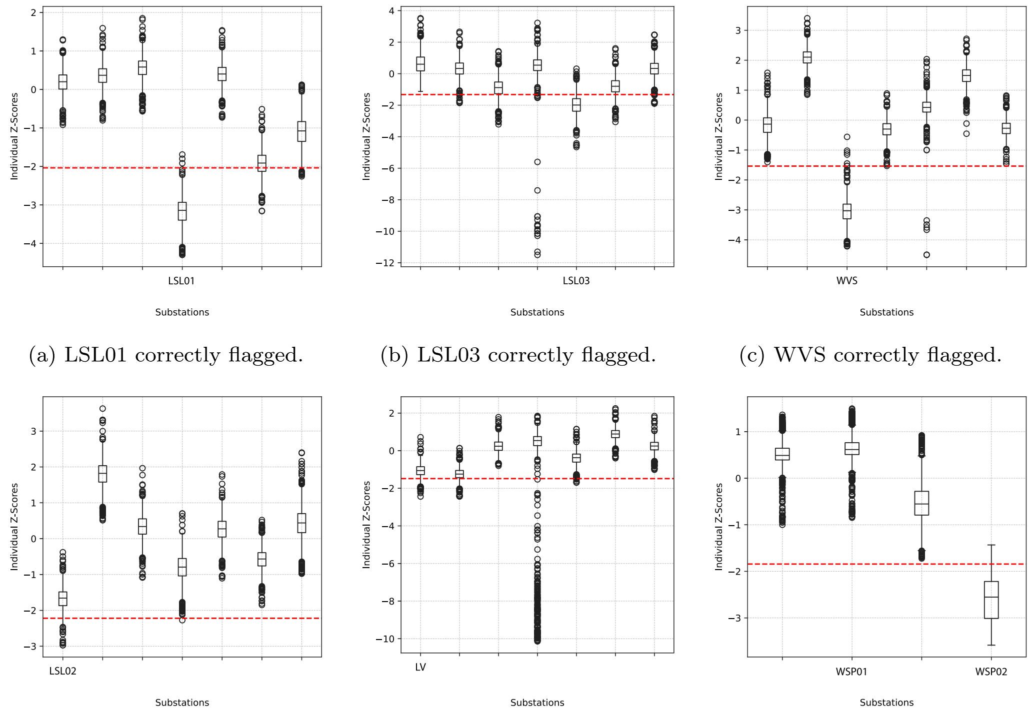
5.2.2. Performance of HEAT fault detection

The proposed HEAT method achieved the best performance, with a sensitivity of 74.1% ± 1.99 and specificity of 95.50% ± 0.31, respectively. These results demonstrate the effectiveness of leveraging small and meaningful substation clusters for fault detection. HEAT not only surpassed the arbitrary equal cluster baseline in sensitivity ($T = 7.51$; $p < 0.001$; $d = 3.36$) and specificity ($T = 3.45$; $p < 0.001$; $d = 1.55$), it is also statistically significant with a large effect size compared to [29], which reported a sensitivity of 65.00% ($T = 9.16$; $p < 0.001$; $d = 2.83$). While the accuracy provides an approximate, it is limited due to a share of unlabelled substations. In some cases, unlabelled substations have been flagged as faulty, but independent verification of these substations is not feasible. However, analysis suggest that these substations exhibit behaviours similar to those of confirmed faulty substations, implying that the true performance of the method may be higher than reported, our evaluation thus represents a conservative estimate.

The effect sizes observed, particularly the substantial improvement in sensitivity when transitioning from global detection to HEAT, have significant practical implications for DH systems. Higher sensitivity directly translates into a greater likelihood of early identification of faulty substations, reducing the risk of prolonged inefficiencies or undetected failures that could escalate maintenance costs or impact heat delivery. By forming meaningful small substation groups for localised detection, HEAT enhances the detection of both pronounced (e.g., wrong valve settings Fig. 6(c) or sensor issues Fig. 6(f)) and subtle faults (e.g., secondary leakages Figs. 6(a) and 6(b)), achieving high sensitivity and specificity. While global detection provides a baseline, its inability to handle subtle variations across substations renders it ineffective for practical applications. Random clustering mitigates some of these issues by reducing noise within smaller groups, but its arbitrary nature limits further improvements. While HEAT generally improved performance, it also misses some faults as illustrated in Figs. 6(d), 6(e), or 6(f), albeit obfuscated by more pronounced faults; suggesting room for improvement.

As seen in Fig. 6, applying MAD z-scores results in a reduced spread in the corresponding box plot with more outliers. This narrower distribution results in fewer substations being flagged as faulty, as it becomes harder to surpass 2σ, which enhances the specificity of our detection process by reducing the false positive rate. However, this increased specificity comes with the trade-off of potentially overlooking some genuinely faulty substations, i.e., MAD z-scores improve robustness by mitigating the influence of extreme values; they may also reduce sensitivity. For instance, Fig. 6(d) shows that the large secondary valve leak is missed because the substation is not considered 2σ away from the intra-cluster distribution. This could result from high intra-cluster variability or an unknown faulty substation skewing the distribution. Moreover, natural heterogeneity among substations, such as a predominance of unknown faults, might render a uniform threshold less effective. While lowering the threshold could improve fault detection, it would likely increase the false positive rate, underscoring the trade-off between sensitivity and specificity and the potential benefit of adaptive thresholding based on local distribution characteristics. Another example is provided in Fig. 6(e), where a large valve fault is not detected. Here, a single substation with numerous outliers skews the overall distribution, diminishing the influence of other poorly performing substations and reducing fault detection likelihood. Despite these misses, the final results can still inform DH operators, allowing further investigation and targeted maintenance. Notably, none of the labelled 'normal' substations were flagged as faulty, indicating that HEAT effectively minimises false positives and thereby avoids unnecessary maintenance interventions while optimising resource allocation. Furthermore, the method's ability to operate without labelled data is particularly advantageous in the DH industry, where such datasets are extremely scarce [12,17]. By directing maintenance resources towards substations with the highest potential for faults, HEAT can facilitate the creation of high-quality ground-truth datasets. This iterative strategy, where maintenance interventions recalibrate the MAD z-scores and reveal new underperforming substations has the potential to enhance operational efficiency and progressively improve ML model performance and advance intelligent FDD in the DH industry.

HEAT demonstrated robust performance on the Southern Shandong network: statistical evaluations (Welch's t-test, Cohen's d, and confidence intervals) confirm that its improvements are not due to random chance but are systematically driven by the data and soft-constraint structure. Given that the thermodynamic relationships governing heat loss and flow in DH networks are universal, we anticipate that HEAT will generalise well to other systems under similar operating regimes. Nonetheless, applying HEAT to networks with different configurations, such as looped or mesh topologies, larger spatial scales, or varied sampling rates, may introduce new challenges and remain an open question. HEAT's network-agnostic design (normalised pairwise distances in $[0,1]$, penalty weights, and cluster-size bounds defined as



(a) LSL01 correctly flagged. (b) LSL03 correctly flagged. (c) WVS correctly flagged.

(d) LSL02 not flagged due to heterogeneity among substations. (e) LV not flagged; obscured due to many outliers in substation 4. (f) WSP2 correctly flagged but obscures another WSP1 fault.

Fig. 6. Each sub-figure represents an individual cluster, with each box plot representing a substation. The red dotted line represents the threshold of 2σ. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

simple functions of the total number of substations) enables the transfer of the soft-constraint framework with minimal recalibration. However, substantial differences in scale and design may require modified constraint formulations, and data collected at different frequencies would necessitate retraining the CAE and adjusting its latent space. We identify these adaptations as key topics for future work to fully assess HEAT's adaptability and limitations across diverse DH contexts.

5.3. Limitations

Despite the promising performance of HEAT, several limitations should be considered when interpreting the results. These constraints arise from the nature of the data and the methodological framework, which may affect the true accuracy and generalisability. The following list summarises the primary limitations of this study:

- **Partial Labelling:** The dataset used in this study is partially labelled, with annotations provided by an independent domain expert and the DH utility. Although these labels are valuable for validation, they may overlook some faulty substations. Thus, the true accuracy of HEAT might be affected or skewed by the incompleteness of the provided labels.
- **Outlier Dominance:** The statistical model is designed to identify the worst-performing substations. However, when extremely poor-performing substations dominate the distribution, they may mask less severe but still significant faults. This limitation necessitates re-applying HEAT after maintenance interventions to reveal additional issues.
- **Parameter Sensitivity:** The performance of HEAT is sensitive to the tuning of parameters (e.g., thresholds for outlier detection or clustering parameters). Inaccurate calibration could lead to a trade-off between missing potential faults (reduced sensitivity) and increasing false positives (reduced specificity).
- **Scalability and Generalisability:** Although HEAT performs well in the context of the limited number of DH substations analysed here, its scalability to larger systems and generalisability across different types of networks require further investigation. Nonetheless, the fundamental thermodynamic principles governing DH systems suggest that HEAT is likely to generalise effectively.

6. Conclusion

HEAT is a Hierarchical-constrained Encoder-Assisted Time Series clustering method specifically developed to enhance fault detection in DH substations; by integrating constrained clustering techniques with domain-specific knowledge, HEAT forms meaningful substation groupings that balance cluster cohesion with uniformity. Its linkage function embeds adaptive soft penalty constraints, enabling the enforcement of maximum and minimum cluster sizes and must-link and cannot-link constraints. This informed, flexible approach guides the clustering process beyond traditional global assessments, leveraging the cluster solution for relative comparisons within substation groups.

Experimental results show that HEAT significantly outperforms conventional clustering and global detection methods, achieving a sensitivity of 74.10% and a specificity of 95.50%. Moreover, operating in an unsupervised mode without needing labelled data offers a practical, scalable solution for real-world DH applications; improving operational efficiency, reducing maintenance costs, and enhancing energy sustainability in DH networks.

Future research will focus on several key areas to enhance the effectiveness and applicability of HEAT. First, we plan to explore alternative adaptive thresholding strategies to complement our current intra-cluster 2σ rule and potentially enhance fault-detection accuracy. Second, further refinements to the CAE architecture may enhance the latent space's ability to discriminate between dissimilar samples, thereby yielding more robust clustering results. Third, evaluate HEAT on DH systems with multiple independent heat plants, providing a complex setting. This will test constraint transferability and network-agnostic parameter settings and identify the scenarios in which HEAT's topology-aware, dendrogram-based clustering outperforms industry standard methods (e.g. k-means. Finally, collaborative efforts with DH utility providers could facilitate the creation of a more comprehensive labelled dataset, improving model training and validation.

CRediT authorship contribution statement

Jonne van Dreven: Visualization, Software, Formal analysis, Writing – review & editing, Validation, Investigation, Conceptualization, Writing – original draft, Methodology. **Abbas Cheddad:** Writing – review & editing, Supervision, Investigation, Visualization, Methodology, Writing – original draft, Project administration, Formal analysis. **Ahmad Nauman Ghazi:** Writing – original draft, Project administration, Investigation, Visualization, Methodology, Formal analysis, Writing – review & editing, Supervision. **Sadi Alawadi:** Supervision, Writing – original draft, Project administration, Investigation, Writing – review & editing, Visualization, Methodology, Formal analysis. **Jad Al Koussa:** Writing – original draft, Project administration, Data curation, Supervision, Formal analysis, Writing – review & editing, Resources, Investigation. **Dirk Vanhoudt:** Writing – review & editing, Resources, Formal analysis, Supervision, Project administration, Data curation, Writing – original draft, Investigation.

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.

Acknowledgements

We acknowledge and thank *Runa Smart Equipment* Co. Ltd. (China) for providing the district heating data essential for this research. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The data that has been used is confidential.

References

[1] IRENA. Power to heat and cooling: Status. 2022, https://www.irena.org/Innovation-landscape-for-smart-electrification/Power-to-heat-and-cooling/Status, [Accessed 27 November 2024].
[2] Li X, Zhao L, Qin Y, Oleson K, Zhang Y. Elevated urban energy risks due to climate-driven biophysical feedbacks. Nat Clim Chang 2024;14(10):1056–63. http://dx.doi.org/10.1038/s41558-024-02108-w.
[3] Marszal-Pomianowska A, Motoasca E, Pothof I, Felsmann C, Heiselberg P, Cadenbach A, Leusbrock I, O'Donovan K, Petersen S, Schaffer M. Strengths, weaknesses, opportunities and threats of demand response in district heating and cooling systems. from passive customers to valuable assets. Smart Energy 2024;14:100135. http://dx.doi.org/10.1016/j.segy.2024.100135.
[4] Werner S. International review of district heating and cooling. Energy 2017;137:617–31. http://dx.doi.org/10.1016/j.energy.2017.04.045.
[5] IEA. 350 million building units connected to district energy networks by 2030, provide about 20% of space heating needs. 2022, URL Link, [Accessed 01 August 2023].
[6] Lund H, Østergaard PA, Chang M, Werner S, Svendsen S. The status of 4th generation district heating: Research and results. Energy 2018;164:147–59. http://dx.doi.org/10.1016/j.energy.2018.08.206.
[7] Lund H, Thorsen JE, Jensen SS, Madsen FP. Fourth-generation district heating and motivation tariffs. ASME Open J Eng 2022;1(011002). http://dx.doi.org/10.1115/1.4053420.
[8] Månsson S, et al. Automated statistical methods for fault detection in district heating customer installations. Energies 2019;12(11):113. http://dx.doi.org/10.3390/en12010113.
[9] Gadd H, Werner S. Fault detection in district heating substations. Appl Energy 2015;157:51–9. http://dx.doi.org/10.1016/j.apenergy.2015.07.061.
[10] Honoré K. The age of digitalization and flexibility - from consumer to flexumer in the district heating system. In: 9th international conference on smart energy systems. Copenhagen; 2023.
[11] Schmidt D, et al., editors. Guidebook for the digitalisation of district heating: Transforming heat networks for a sustainable future. Germany: AGFW Project Company, Frankfurt am Main; 2023, p. 67, Final Report of DHC Annex TS4.
[12] van Dreven J, Boeva V, Abghari S, Grahn H, Al Koussa J, Motoasca E. Intelligent approaches to fault detection and diagnosis in district heating: Current trends, challenges, and opportunities. Electron 2023;12(66):1448. http://dx.doi.org/10.3390/electronics12061448.
[13] Rafati A, Shaker HR. Predictive maintenance of district heating networks: A comprehensive review of methods and challenges. Therm Sci Eng Prog 2024;53:102722. http://dx.doi.org/10.1016/j.tsep.2024.102722.
[14] Guo H, Huang B, Wang J. Probabilistic load forecasting for integrated energy systems using attentive quantile regression temporal convolutional network. Adv Appl Energy 2024;14:100165. http://dx.doi.org/10.1016/j.adapen.2024.100165.
[15] Frederiksen S, Werner S. District heating and cooling. 2013.
[16] Leiria D, et al. Towards automated fault detection and diagnosis in district heating customers: generation and analysis of a labeled dataset with ground truth. In: Building simulation. vol. 18, IBPSA; 2023, p. 3620–8. http://dx.doi.org/10.26868/25222708.2023.1576, URL https://publications.ibpsa.org/conference/paper/?id+bs2023_1576.
[17] 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 2023;276:127569. http://dx.doi.org/10.1016/j.energy.2023.127569.
[18] Gadd H, Werner S. Achieving low return temperatures from district heating substations. Appl Energy 2014;136:59–67. http://dx.doi.org/10.1016/j.apenergy.2014.09.022.
[19] Månsson S, et al. A machine learning approach to fault detection in district heating substations. Energy Procedia 2018;149:226–35. http://dx.doi.org/10.1016/j.egypro.2018.08.187.
[20] Jangsten M, Lindholm T, Dalenbäck J-O. A performance assessment method for district cooling substations based on operational data. Sci Technol Built Environ 2022;28(10):1472–88. http://dx.doi.org/10.1080/23744731.2022.2129272.
[21] T F, et al. Fault detection and condition monitoring in district heating using smart meter data. PHM Soc Eur Conf 2021;6(1):11. http://dx.doi.org/10.36001/phme.2021.v6i1.2786.
[22] Calikus E, et al. Ranking abnormal substations by power signature dispersion. Energy Procedia 2018;149:345–53.
[23] Gianniou P, et al. Clustering-based analysis for residential district heating data. Energy Convers Manage 2018;165:840–50.
[24] Tureczek AM, et al. Clustering district heat exchange stations using smart meter consumption data. Energy Build 2019;182:144–58.
[25] Hong Y, Yoon S. Holistic operational signatures for an energy-efficient district heating substation in buildings. Energy 2022;250:123798.
[26] Xue P, et al. Fault detection and operation optimization in district heating substations based on data mining techniques. Appl Energy 2017;205:926–40.
[27] Koussa JA, Månsson S. Fault detection in district heating substations: a cluster-based and an instance-based approach. 2022.
[28] Calikus E, et al. A data-driven approach for discovering heat load patterns in district heating. Appl Energy 2019;252:113409. http://dx.doi.org/10.1016/j.apenergy.2019.113409.
[29] Van Dreven J, Cheddad A, Alawadi S, Ghazi AN, Al Koussa J, Vanhoudt D. Shedad: Snn-enhanced district heating anomaly detection for urban substations. In: 2024 9th international conference on fog and mobile edge computing. FMEC, 2024, p. 130–7. http://dx.doi.org/10.1109/FMEC62297.2024.10710205, URL https://ieeexplore.ieee.org/document/10710205.
[30] Bradley PS, Bennett KP, Demiriz A. Constrained k-means clustering, vol. 20, Microsoft Research, Redmond; 2000.
[31] Wagstaff K, Cardie C, Rogers S, Schrödl S, et al. Constrained k-means clustering with background knowledge. In: Icml. vol. 1, 2001, p. 577–84.
[32] Wagstaff K, Cardie C. Clustering with instance-level constraints. In: Proceedings of the seventeenth international conference on machine learning. ICML '00, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2000, p. 1103–10.
[33] Bar-Hillel A, Hertz T, Shental N, Weinshall D. Learning distance functions using equivalence relations, ICML'03. Washington, DC, USA: AAAI Press; 2003, p. 11–8.
[34] Xing EP, Ng AY, Jordan MI, Russell S. Distance metric learning, with application to clustering with side-information. In: Proceedings of the 15th international conference on neural information processing systems. NIPS'02, Cambridge, MA, USA: MIT Press; 2002, p. 521–8.
[35] Basu S, Banerjee A, Mooney RJ. Active semi-supervision for pairwise constrained clustering. Proc Soc Ind Appl Math 2004;333–44. http://dx.doi.org/10.1137/1.9781611972740.31, URL https://epubs.siam.org/doi/10.1137/1.9781611972740.31.
[36] Davidson I, Ravi SS. Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. In: Jorge AM, Torgo L, Brazdil P, Camacho R, Gama J, editors. Knowledge discovery in databases: PKDD 2005. Berlin, Heidelberg: Springer; 2005, p. 59–70. http://dx.doi.org/10.1007/11564126_11.

[37] Davidson I, Ravi SS. Using instance-level constraints in agglomerative hierarchical clustering: theoretical and empirical results. Data Min Knowl Discov 2009;18(2):257–82. http://dx.doi.org/10.1007/s10618-008-0103-4.

[38] Chong CY, Lee SP. Constrained agglomerative hierarchical software clustering with hard and soft constraints. In: 2015 international conference on evaluation of novel approaches to software engineering. ENASE, 2015, p. 177–88, URL https://ieeexplore.ieee.org/document/7320352.

[39] Ambroise C, Dehman A, Neuvial P, Rigaill G, Vialaneix N. Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics. Algorithms Mol Biol 2019;14(1):22. http://dx.doi.org/10.1186/s13015-019-0157-4.

[40] Mauduit E, Simonetto A. Constrained hierarchical clustering via graph coarsening and optimal cuts. In: 2023 57th asilomar conference on signals, systems, and computers. 2023, p. 844–8. http://dx.doi.org/10.1109/IEEECONF59524.2023.10476840, URL https://ieeexplore.ieee.org/document/10476840.

[41] Zhao J, Shan Y. An influencing parameters analysis of district heating network time delays based on the cfd method. Energies 2019;12(77):1297. http://dx.doi.org/10.3390/en12071297.

[42] Aggarwal CC, Hinneburg A, Keim DA. On the surprising behavior of distance metrics in high dimensional space. In: International conference on database theory. Springer; 2001, p. 420–34.

[43] Wang X, Smith K, Hyndman R. Characteristic-based clustering for time series data. Data Min Knowl Discov 2006;13(3):335–64. http://dx.doi.org/10.1007/s10618-005-0039-x.

[44] Lafabregue B, Weber J, Gançarski P, Forestier G. End-to-end deep representation learning for time series clustering: a comparative study. Data Min Knowl Discov 2022;36(1):29–81. http://dx.doi.org/10.1007/s10618-021-00796-y.

[45] Alawadi S, Fernández-Delgado M, Mera D, Barro S. Polynomial kernel discriminant analysis for 2d visualization of classification problems. Neural Comput Appl 2019;31:3515–31.

[46] de Oliveira MLL, Bekooij MJG. Deep convolutional autoencoder applied for noise reduction in range-doppler maps of fmcw radars. In: 2020 IEEE international radar conference. RADAR, 2020, p. 630–5. http://dx.doi.org/10.1109/RADAR42522.2020.9114719, URL https://ieeexplore.ieee.org/document/9114719.

[47] Campos D, Kieu T, Guo C, Huang F, Zheng K, Yang B, Jensen CS. Unsupervised time series outlier detection with diversity-driven convolutional ensembles. Proc VLDB Endow 2021;15(3):611–23. http://dx.doi.org/10.14778/3494124.3494142.

[48] Milani AE, Zappalá D, Watson SJ. A hybrid convolutional autoencoder training algorithm for unsupervised bearing health indicator construction. Eng Appl Artif Intell 2025;139:109477. http://dx.doi.org/10.1016/j.engappai.2024.109477.

[49] Bourlard H, Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition. Biol Cybernet 1988;59(4):291–4. http://dx.doi.org/10.1007/BF00332918.

[50] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Sci 2006;313(5786):504–7. http://dx.doi.org/10.1126/science.1127647.

[51] Liu T, Wang J, Liu Q, Alibhai S, Lu T, He X. High-ratio lossy compression: Exploring the autoencoder to compress scientific data. IEEE Trans Big Data 2023;9(1):22–36. http://dx.doi.org/10.1109/TBDATA.2021.3066151.

[52] Ahmed T, Longo L. Examining the size of the latent space of convolutional variational autoencoders trained with spectral topographic maps of eeg frequency bands. IEEE Access 2022;10:107575–86. http://dx.doi.org/10.1109/ACCESS.2022.3212777.

[53] Crosby T. How to detect and handle outliers. Technometrics 1994. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1994.10485810.

[54] van Dreven J, Boeva V, Abghari S, Grahn H, Al Koussa J. A systematic approach for data generation for intelligent fault detection and diagnosis in district heating. Energy 2024;307:132711. http://dx.doi.org/10.1016/j.energy.2024.132711.

[55] Leiria D, Johra H, Anoruo J, Praulins I, Piscitelli MS, Capozzoli A, Marszal-Pomianowska A, Pomianowski MZ. Is it returning too hot? time series segmentation and feature clustering of end-user substation faults in district heating systems. Appl Energy 2025;381:125122. http://dx.doi.org/10.1016/j.apenergy.2024.125122.

[56] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. Nature Methods 2020;17(3):261–72.

[57] Tavenard R, Faouzi J, Vandewiele G, Divo F, Androz G, Holtz C, Payne M, Yurchak R, Rußwurm M, Kolar K, et al. Tslearn, a machine learning toolkit for time series data. J Mach Learn Res 2020;21(118):1–6.

[58] Abadi M, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. 2015, software available from tensorflow.org URL https://www.tensorflow.org/.

[59] Ward Jr JH. Hierarchical grouping to optimize an objective function. J Amer Statist Assoc 1963;58(301):236–44. http://dx.doi.org/10.1080/01621459.1963.10500845.

[60] MacQueen J. Some methods for classification and analysis of multivariate observations. vol. 5.1, University of California Press; 1967, p. 281–98.

[61] Hartigan JA, Wong MA. Algorithm as 136: A k-means clustering algorithm. J R Stat Soc Ser C (Appl Stat) 1979;28(1):100–8. http://dx.doi.org/10.2307/2346830.

[62] Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the second international conference on knowledge discovery and data mining. KDD'96, Portland, Oregon: AAAI Press; 1996, p. 226–31.

[63] McInnes L, Healy J, Astels S. Hdbscan: Hierarchical density based clustering. J Open Source Softw 2017;2(11):205. http://dx.doi.org/10.21105/joss.00205.

[64] Hinneburg A, Gabriel H-H. Denclue 2.0: Fast clustering based on kernel density estimation. In: International symposium on intelligent data analysis. Springer; 2007, p. 70–80.

[65] Ng AY, Jordan MI, Weiss Y. On spectral clustering: analysis and an algorithm. In: Proceedings of the 15th international conference on neural information processing systems: Natural and synthetic. NIPS'01, Cambridge, MA, USA: MIT Press; 2001, p. 849–56.

[66] Paparrizos J, Gravano L. K-shape: Efficient and accurate clustering of time series. SIGMOD Rec 2016;45(1):69–76. http://dx.doi.org/10.1145/2949741.2949758.

[67] Overall JE, Atlas RS, Gibson JM. Power of a test that is robust against variance heterogeneity. Psychol Rep 1995;77(1):155–9. http://dx.doi.org/10.2466/pr0.1995.77.1.155.

[68] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed.. New York: Routledge; 2013, http://dx.doi.org/10.4324/9780203771587.