Postprint

Citation for the original published paper:

van Dreven, J., Cheddad, A., Alawadi, S., Ghazi, A N., Al Koussa, J. et al. (2024)
SHEDAD: SNN-Enhanced District Heating Anomaly Detection for Urban Substations
In: Quwaider M., Alawadi S., Jararweh Y. (ed.), *2024 9th International Conference on
Fog and Mobile Edge Computing, FMEC 2024* (pp. 130-137). Institute of Electrical and
Electronics Engineers (IEEE)
https://doi.org/10.1109/FMEC62297.2024.10710205

N.B. When citing this work, cite the original published paper.

SHEDAD: SNN-Enhanced District Heating Anomaly Detection for Urban Substations

Jonne van Dreven*‡§, Abbas Cheddad*, Sadi Alawadi*, Ahmad Nauman Ghazi†, Jad Al Koussa‡§, Dirk Vanhoudt‡§

* Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden
† Department of Software Engineering, Blekinge Institute of Technology, Karlskrona, Sweden
‡ Unit Energy Technology, Flemish Institute for Technological Research (VITO), Mol, Belgium
§ EnergyVille, Genk, Belgium
jonne.van.dreven@bth.se

Abstract—District Heating (DH) systems are essential for energy-efficient urban heating. However, despite the advancements in automated fault detection and diagnosis (FDD), DH still faces challenges in operational faults that impact efficiency. This study introduces the Shared Nearest Neighbor Enhanced District Heating Anomaly Detection (SHEDAD) approach, designed to approximate the DH network topology and allow for local anomaly detection without disclosing sensitive information, such as substation locations. The approach leverages a multi-adaptive k-Nearest Neighbor (k-NN) graph to improve the initial neighborhood creation. Moreover, it introduces a merging technique that reduces noise and eliminates trivial edges. We use the Median Absolute Deviation (MAD) and modified z-scores to flag anomalous substations. The results reveal that SHEDAD outperforms traditional clustering methods, achieving significantly lower intra-cluster variance and distance. Additionally, SHEDAD effectively isolates and identifies two distinct categories of anomalies: supply temperatures and substation performance. We identified 30 anomalous substations and reached a sensitivity of approximately 65% and specificity of approximately 97%. By focusing on this subset of poor-performing substations in the network, SHEDAD enables more targeted and effective maintenance interventions, which can reduce energy usage while optimizing network performance.

Index Terms—Anomaly Detection, Clustering, District Heating, Nearest Neighbor Measure, Intelligent Urban Systems

I. INTRODUCTION

The high energy consumed by cities worldwide for heating purposes significantly impacts climate change, the surrounding environment and the economy. Essential actions and solutions must be considered to face this challenge. For instance, district Heating (DH) systems offer a solution to meet urban heating demands while leveraging renewable energies efficiently. DH systems generate heat centrally and then distribute it via a network of insulated pipes, making them one of the most sustainable methods for providing heat in densely populated areas [26]. The International Energy Agency (IEA) expects that by 2030, approximately 350 million buildings will be connected to DH systems, fulfilling 20% of global space heating requirements [8]. Despite their benefits, unlocking the immense possibilities of DH systems will require carefully addressing the substantial challenges that presently constrain their capabilities.

One major challenge is the presence of operational faults in DH substations, which negatively impacts the substation itself and the entire network. Studies indicate that a substantial percentage (approximately 43 and 75% of the substations) of DH substations can operate sub-optimal due to faults [5], [16]. The lack of automated Fault Detection and Diagnosis (FDD) methods aggravates this issue, allowing faults to persist for extended periods, significantly impacting energy use. Fortunately, with the introduction of automatic heat meters, DH networks become better equipped for data-driven solutions and, with the proper detection, could help reduce the DH customers' energy usage on average by 14% [13]3. Although, current data collection is not standardized and is primarily focused on billing rather than FDD. Commonly, DH datasets comprise primary side hourly measurements, including features such as supply temperature, return temperature, and flow rate, alongside derived metrics such as energy consumption. Most often, secondary side information (supply/return temperatures, flow, set points, and indoor temperature) and supplementary information, such as layouts of DH networks and geographical substation locations, are confidential, limiting the depth and spatial analysis that can be conducted. Additionally, DH data contains much-introduced noise in the measurements, e.g., supply temperature measurements contain substation behavior due to the sensor placement, as well as unknown confounders influencing measurements, e.g., building occupancy, size, or domestic hot water usages, which makes many conventional analytical methods often fall short. Furthermore, the lack of labeled data [18] amplifies the challenges, making it difficult to develop Machine Learning (ML) methods and validate them. Given the complexity and constraints of DH data, there is a need for tailored methods specifically for DH.

Automatic FDD is commonly done through three steps. The process begins with *fault detection*, where algorithms monitor system parameters to identify deviations. This is followed by *fault diagnosis*, which involves analyzing the data to pinpoint the fault's cause. The final step, *fault correction*, adjusts system settings automatically or prompts manual interventions to remedy the identified issues, thereby ensuring system integrity and operational continuity.

This study introduces an innovative approach, the *Shared*

Nearest Neighbor Enhanced District Heating Anomaly Detection (SHEDAD), designed to approximate the relative topology of DH networks. By utilizing local operational data and identifying similarities, SHEDAD significantly enhances anomaly detection precision. This study is inspired by prior investigations of [11], [14], [20], [27].

We use the consumer supply temperature profiles as a distance metric to define relative neighbors of DH substations, thereby allowing for localized anomaly detection (intra-cluster) using modified z-scores. We identify two distinct categories of anomalies: those related to supply temperatures and those concerning substation performance. Crucially, our method avoids the need for sensitive data, such as geographical substation locations and network layouts, effectively addressing privacy and security concerns while enabling comparative assessments of substation performance. We advance the traditional Shared Nearest Neighbor (SNN) clustering method by integrating several enhancements tailored for time series DH data. We focus on improving the initial neighborhood creation through a multi-adaptive k-NN graph, which helps to isolate anomalous substations while improving connectivity for regular substations. Additionally, we introduce a novel merging technique that significantly reduces noise and eliminates meaningless edges that often obscure traditional graphs, thereby clarifying the graph structure and enhancing the effectiveness of the clustering process.

II. RELATED WORK

Anomaly detection in DH systems typically involves either a global comparison of all substations against each other across the entire system to identify unusual patterns [15], [16] or using regression analysis on individual substations to detect anomalies based on fixed thresholds [1], [22]. Regression analysis can detect sudden variations in substation behavior, i.e., when a fault occurs. However, an anomaly may remain undetected if the model is constructed during an existing fault. Local anomaly detection is frequently overlooked in many settings due to the constraints imposed by the confidentiality of geographical substation locations. However, such localized analysis can be particularly insightful for FDD. Comparing the performance of substations that experience similar supply temperatures offers a fairer assessment than comparing those with differing temperatures. Furthermore, understanding the time delays in supply temperature across the network allows for strategic adjustments in flow rates, which can optimize and potentially reduce these delays. Such optimization ensures more efficient heat distribution and enhances the overall operational efficiency of the DH system.

Clustering, such as k-means and k-shape, has been widely explored in DH, for instance, to analyze substation heat consumption patterns [6], [7], [23], outlier detection [12], [28], or discover heat load patterns [2]. While k-means is effective, it is sensitive to noise and outliers and cannot handle non-globular patterns, limiting its performance in complex datasets, such as in DH. Therefore, SNN [10] could offer advantages when handling DH data, as it focuses on the density of shared nearest neighbors rather than centroid-based distance. This method effectively reduces the impact of outliers and noise, enhancing the cluster results in complex datasets. Recent advancements in SNN have significantly improved the precision and clarity of identifying cluster boundaries. Kim et. al. [11] proposed a SNN-based method to identify the fault variables of a circulating fluidized bed boiler. They applied it to timely fault detection and isolation (FDI). In [27], a hierarchical clustering-based method utilizing structural similarities in nearest neighbor graphs (HCNN) is introduced, which outperforms traditional methods like the Density-based Spatial Clustering of Applications with Noise (DBSCAN), the Density Peak Clustering (DPC), and the k-means algorithms in handling unclear boundaries. Similarly, in [20], a Selective Nearest Neighbors Clustering (SNNC) technique is presented, effectively reducing weak connections and enhancing border and outlier detection, thereby outperforming conventional approaches in complex datasets. Furthermore, in [14], an adaptation of the SNN framework combined with DPC utilizes Jaccard similarity metrics to refine cluster center identification and data point assignment, providing a marked improvement over traditional DPC by incorporating nearest-neighbor information for better cluster allocation and center discovery.

This study addresses the limitations of traditional anomaly detection strategies for DH systems, which typically rely on global comparisons that may overlook local nuances. Our SHEDAD method adopts a localized approach through a relative topology approximation, thereby enhancing the precision of anomaly detection and enabling more targeted maintenance interventions.

III. DATA ACQUISITION

The dataset used in this study comprises sensor measurements linked to different substations and geographical data of a DH network located in the southern part of Shandong province, China. Specifically, the dataset consists of data recorded every 5 minutes from 248 substations. The data collection spanned one month, from 1st January 2024 to 31st January 2024 (8,928 samples), capturing operational measurements of each substation as described in Table I, across a broad spectrum of outdoor temperatures ranging from $-9°C$ to $11°C$. This wide range of temperatures allows for detailed analysis of the substations under varying climatic conditions. Due to the sensitive nature of the information, the dataset is confidential and has not been made publicly available.

TABLE I: Delineation of the Features Comprising the Dataset

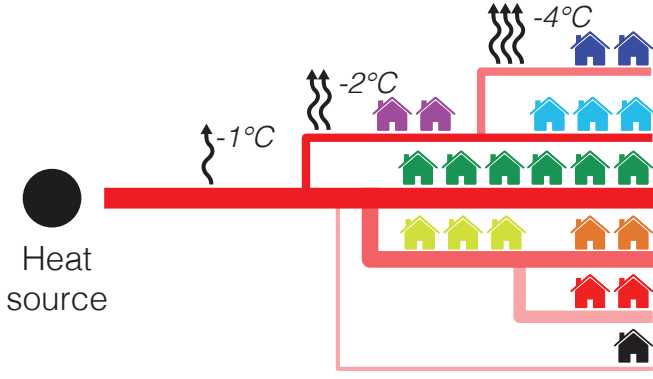| Feature	Type	Description
Timestamp	Datetime	Date and time of measurement
Supply temp.	Continuous	Primary supply temperature
Return temp.	Continuous	Primary return temperature
Flow	Continuous	Primary flow rate
Outdoor Temp.	Continuous	Outdoor temperature
X Location	Continuous	X coordinate
Y location	Continuous	Y coordinate

Fig. 1: Illustration of a DH network with a single heat source and various pipe diameters and flow rates. Substation supply temperatures are affected due to heat loss during transportation. Each color represents a neighboring group of substations with similar supply temperature profiles.

IV. METHOD

Our approach, SHEDAD, aims to systematically identify two distinct types of anomalies within DH networks: supply temperature and performance anomalies.

Initially, we construct a relative network topology, which utilizes supply temperature profiles as a distance metric to approximate spatial relationships and operational similarities between substations. This topology serves as the framework for our anomaly detection processes, enabling us to pinpoint local deviations in supply temperature that may indicate underlying issues. The anomalous substations are separated into singleton clusters, while substations with similar supply temperature characteristics are grouped together.

Secondly, leveraging the established network topology, we implement an anomaly detection strategy focused on substation performance. This involves analyzing each substation's operational performance relative to its immediate topological neighbors. We detail each phase of SHEDAD in the subsequent sections. Below, we describe the processes involved in constructing the network topology in subsection IV-A and the specific techniques used for subsequent anomaly detection based on substation performance in subsection IV-B.

A. Network Topology

The primary supply temperatures reaching the consumer substations are affected by several aspects that contribute to the heat losses [29]. The aspects are:

- Flow rate: The rate at which the heated water or medium is circulated through the network impacts the supply temperature. Higher flow rates can lead to reduced time delays in heat delivery. Flow rate is influenced by a combination of factors related to the heat source, network design, and the consumption patterns of each heat consumer.
- Pipe Length and Diameter: Longer distances and variations in pipe diameter affect the time it takes for the heat

to reach different parts of the network, which can lead to variations in supply temperature at different substations.
- Water Supply Temperature: The initial temperature of the water supplied from the heat source directly influences the supply temperatures throughout the network. Any fluctuations at the source are propagated throughout the system.

Figure 1 illustrates a relative network topology with its respective heat losses. Since the combination of these factors becomes unique for each substation, we can use the supply temperature profile to approximate the relative location of a substation to the heat source and its neighboring substations. As shown in Figure 2, the method has three main steps.

1) preprocessing: The preprocessing step prepares the supply temperature profiles from each substation for subsequent analysis. This phase ensures the quality and consistency of the data, which are critical for the reliability of the analysis. At first, we used standard data cleaning procedures, e.g., removing substations with missing values, which resulted in removing 28 substations.

We start by segmenting the dataset into daily supply profiles. The segmentation allows for studying individual days instead of the entire period, which can help avoid issues that might arise when clustering extensive time series data. For example, significant patterns or anomalies on specific days might get averaged out or diluted if the entire dataset is used.

To ensure our model accurately reflects the diverse operational dynamics of DH networks, it is critical to sample data from a broad spectrum of conditions. For instance, supply temperature variations are influenced by outdoor temperatures, as on colder days, increased flow rates may result in smaller network delays. In contrast, warmer days might see reduced flow and larger network delays. We select r random days from the segmented data to mitigate selection bias and prevent the model from overfitting. This random selection process helps us capture a representative cross-section of operational scenarios, ensuring our analysis remains robust and generalizable across the different environmental conditions.

For each of the r randomly selected days, we create a distance matrix between all pairs of substations based on their supply temperature profile. The distances are calculated using Dynamic Time Warping (DTW), with a Sakoe-Chiba warp constraint [19]. This constraint allows for small time shifts to reflect minor delays in the DH network but prevents overly aggressive realignments that could distort the true relationships. This step is crucial for capturing the dynamic nature of temperature changes over time while ensuring that only meaningful temporal variations are considered to improve the model performance.

2) Initialization: In this step, we aim to construct a representative graph G' where nodes N represent substations, and weighted edges E represent the DTW distance between substations. The primary goal is to optimally balance the number of edges, as an excessive number of edges can obscure vital connections by obscuring the graph. Conversely, too few edges create a sparse graph, potentially overlooking meaning-
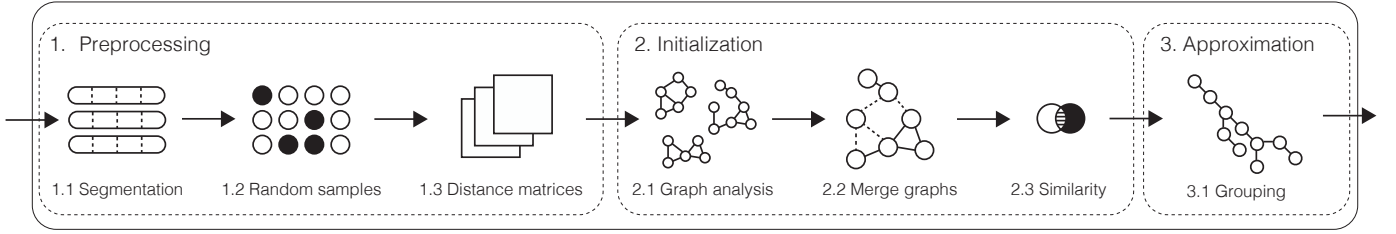
Fig. 2: Diagram of our proposed method for the relative approximation of a DH network.

ful relationships. Our approach aims to fine-tune the number of edges to capture essential connections effectively, ensuring that the graph remains interpretable and comprehensive without being overly complex or overly simplified.

For each distance matrix $d_i \in D$, we construct an adaptive k-nearest neighbor (k-NN) graph G_i. The number of neighbors k for each substation is dynamically adjusted based on its edge weights, allowing more neighbors as long as their weights are below the threshold. Specifically, for substations with high weights, indicating high dissimilarity to other substations, we reduce the associated k value to minimize substation connectivity, effectively isolating them from the main graph. Conversely, nodes with predominantly low-weight connections, indicating closer proximity, are allowed more neighbors, enhancing their connectivity within the graph. This adaptive approach ensures a nuanced representation of substation relationships, balancing the need to maintain overall network cohesion while isolating outlying substations. The criteria for adjusting $k_i = k_b + \Delta k_i$, where k_b is the base number of neighbors, and Δk_i is the adjustment computed as:

$$\Delta k_i = \begin{cases} \left(\frac{\text{low}_i}{k_b} - 1\right) \cdot \frac{k_b}{2}, & \text{if low}_i \geq k_b, \\ \left(\frac{\text{high}_i}{k_b} - 1\right) \cdot \frac{k_b}{2}, & \text{if high}_i \geq k_b, \end{cases} \tag{1}$$

where, for node i, low_i and high_i are the counts of edges with the weight below $\theta_{\min}$ and above $\theta_{\max}$, respectively. Then (i, j) if $\text{dist}(i, j) \leq \theta_{\max}$ and meets the criteria based on Δk_i.

The individual k-NN graphs G_i for $i = 1, 2, \ldots, d$ are merged into a single graph G' using a robust statistical technique namely Fleiss' kappa [4]. Fleiss' kappa measures the extent to which the agreement (inclusion) of a certain edge among the individual graphs exceeds what would be expected by chance. This approach ensures that G' includes only those edges that consistently demonstrate significant agreement across the individual graphs, reducing noise and enhancing the reliability of the merged network effectively.

We calculate a similarity matrix from G' to quantify the similarity between each pair of substations. Typically, this is done using the Jaccard index [9]. However, it solely considers shared neighbors. Therefore, we use a weighted similarity method proposed in [14], which accounts for the number of shared neighbors and the sum of the edge weights connecting

them. The weighted similarity between any two nodes i and j is defined as:

$$S(i, j) = \frac{|\mathcal{N}(i) \cap \mathcal{N}(j)|^2}{\sum_{k \in SNN(i,j)}(w_{ik} + w_{jk})} \tag{2}$$

where $\mathcal{N}(i) \cap \mathcal{N}(j))$ denote the sets of neighbors for nodes i and j, respectively, and w_{ik} and w_{jk} are the weights of the edges connecting these nodes to their shared neighbor k.

3) Approximation: The final step approximates the relative network topology through analysis of the merged graph obtained earlier. We use Agglomerative Hierarchical Clustering [21] and Ward Linkage [25] to analyze the similarity matrix constructed in the previous step. This method starts with each substation as an independent cluster and iteratively merges them based on the similarity measure provided by the SNN matrix. This method creates a dendrogram, a tree-like diagram illustrating how substations merge during the clustering process. This visualization offers clear, interpretable results revealing the substations' hierarchical relationships. Agglomerative hierarchical clustering is particularly suited for DH networks due to its bottom-up approach, which naturally aligns with how substations are interconnected. By starting with the most similar pairs of substations and progressively building larger clusters, this method mirrors the geographic proximities inherent within the network. This incremental clustering process allows for a nuanced understanding of how local operational characteristics aggregate into broader patterns, providing valuable insights into the complex dynamics of DH systems.

B. Anomaly Detection

The topology approximation forms the essential foundation for our anomaly detection system, making it feasible only with the initial processing in place to enable effective detection. We construct a Minimum Spanning Tree (MST) for each cluster based on pairwise Euclidean distances calculated from the supply temperatures. This MST serves as the framework for our performance comparisons, where each substation in the tree is compared against its k neighboring substations. Given that a substation may be compared multiple times—up to $k-1$ times—we normalize the anomaly scores by the number of comparisons each substation undergoes. A normalized score of 1 denotes a significant anomaly, as it has been flagged in

all its comparisons. In contrast, a score of 0 indicates average performance relative to its peers, i.e., it has never been flagged as an anomaly.

For each comparison, we calculate the Median Absolute Deviation (MAD) to robustly measure the dispersion within each comparison of $k + 1$ substations (intra-cluster). Additionally, we use modified z-scores [3] to identify the outlying substations, i.e., substations with absolute mean z-scores exceeding the threshold of two standard deviations below the median are identified as anomalies. This non-parametric approach provides a robust defense against outliers and ensures a more precise measurement of dispersion compared to standard deviation. It is especially effective in scenarios where the data within clusters may not follow a normal distribution, addressing the challenges posed by the diverse and potentially skewed data characteristics encountered in district heating systems.

To assess substation performance, we use the primary ΔT, often seen as a key performance indicator and reflects the efficiency of heat transfer of a substation. ΔT is the difference between the primary supply temperature and the primary return temperature, with high ΔT values suggesting better performance. However, since our analysis focuses on relative comparisons, a substation is considered poor-performing when its mean modified z-score is below 2σ (single-sided) compared to its peers.

C. Experimental Setup

We evaluate our approach against common time-series clustering methods in DH, as outlined in [24]. Specifically, we use time series k-means, Spectral clustering, k-Shape, and SNN. To evaluate the compactness and connectivity of our clusters, we employ the mean MST intra-cluster distance metric, mean intra-cluster variance, and empirical observations. The mean MST intra-cluster distance is defined as:

$$\mathrm{MI} = \frac{1}{n} \sum_{i=1}^{n} d_i, \qquad (3)$$

where MI is the mean intra-cluster distance, d_i denotes the distance between connected nodes within the cluster, and n is the total number of nodes.

The mean intra-cluster variance is defined as:

$$\mathrm{MV} = \frac{1}{n} \sum_{i=1}^{n} (d_i - \overline{d})^2, \qquad (4)$$

where MV is the mean intra-cluster variance, d_i represents the distances between each pair of nodes within a cluster, and $\overline{d}$ is the mean distance calculated from these d_i values. This metric quantifies the variability of distances within a fully connected subgraph of a cluster. A lower variance indicates a high consistency in node connectivity, suggesting that the substations are closely and uniformly integrated. In contrast, a high variance may highlight significant disparities in distances. While our proposed solution is applied to the supply temperature profiles, we validate our outcomes using

the actual geographical substation locations. Such data is typically confidential and, therefore, not commonly accessible. However, their availability in this study allows us to rigorously assess our method's effectiveness in approximating the relative substation locations.

We compute the sensitivity and specificity of our anomaly detection model based on manual analysis provided by a domain expert. Sensitivity, or true positive rate, measures the proportion of actual positives the model correctly identifies. Specifically, it is defined as:

$$\mathrm{sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \qquad (5)$$

On the other hand, specificity, or the true negative rate, measures the proportion of actual negatives that are correctly identified and is given by:

$$\mathrm{specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \qquad (6)$$

It is crucial to consider that the expert analysis, while invaluable, may have its own limitations and not encompass the full spectrum of anomalies. This potential oversight could influence the perceived accuracy of the sensitivity and specificity metrics.

V. RESULTS AND DISCUSSION

This section comprises two subsections, subsection V-A draw our findings on the relative topology approximation and identification of the supply temperature anomalies. At the same time, subsection V-B discusses anomaly detection performance and provides relative comparisons.

A. Relative Topology Approximation

Our study evaluated the performance of multiple clustering algorithms compared to our proposed method on DH data. The evaluation criteria included mean intra-cluster variance and MST distances. As shown in Figure3, SHEDAD was consistently outperforming other clustering methods (from $k > 13$) and maintained a low mean intra-cluster variance (Figure 3a) and distance (Figure 3b) across various cluster sizes k. It is worth highlighting that the proposed approach can recognize both the degree of homogeneity and stability in cluster composition, indicating that the substations within each cluster are closely connected in proximity.

Specifically, the default SNN method demonstrated poor performance, consistently showing high mean intra-cluster variance and MST distances across various cluster sizes. This indicates its inability to capture densely connected clusters effectively. Additionally, it was not stable in finding outlying substations, which affected performance metrics.

k-Means displayed moderate performance, with low mean intra-cluster MST distance scores at $k = 8$ and $k = 10$. However, these scores were misleading, as visual validation indicated poorly formed uniform clusters, with one large cluster and many singletons resulting in a low score. The method exhibited considerable variability across different k
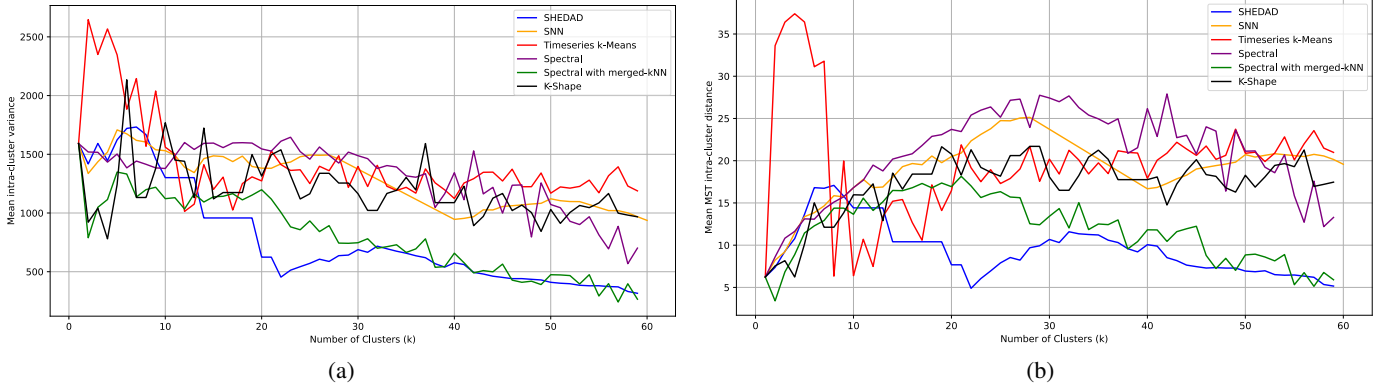
Fig. 3: Performance evaluation for various time-series clustering methods, with mean intra-cluster variance in (a) and mean DTW intra-cluster distance in (b).

and increased scores as the number of k grew. Typically, the performance should increase, as seen in our method, as the number of k grows. k-shape consistently showed a high mean intra-cluster variance and MST distances, indicating poor clustering formations. While k-shape may help to monitor network delays, as it focuses on the shape of data, these network delays do not translate well into a relative topology approximation; substations clusters showed consistent dispersion and non-uniformity. Noteworthy is spectral clustering. While initially, spectral clustering was among the least effective methods, a significant improvement was observed when applying our initialization process (adaptive k-NN graphs merged into a single graph). Notably, for larger values of k (greater than 30), our findings indicate that our enhanced spectral clustering and the proposed SHEDAD method exhibit congruent performance trends. The increase in performance indicates our initialization approach's broad applicability and effectiveness across different clustering algorithms.

In Figure 4, we present several clusters that highlight the impact of noise, which varies from sudden spikes in temperature readings (clusters 2, 5, 6, and 14) or short graduate loss of the supply temperature due to lack of heat demand (cluster 1). The SHEDAD method demonstrated a high degree of resistance to noise, effectively forming dense clusters despite the presence of disturbances, and highlights the capability of our approach to maintain cluster integrity under these noisy conditions.

Additionally, we could effectively separate temperature anomalies into their own clusters. By dynamically adjusting the neighborhood size based on the density of connections (low-weight and high-weight thresholds), our approach separates these outlying components early in the initialization process. This adaptability is crucial for effectively isolating outlier substations that deviate from common patterns. These substations show substantial deviations from most operational patterns and thus are assigned a singleton cluster. The patterns vary from significant fluctuations, unusual behavior, missing readings, or sudden increases. Figure 5 displays a subset of the total 16 supply temperature anomalies we found, showcasing only notable examples. Out of the total 16 supply temperature
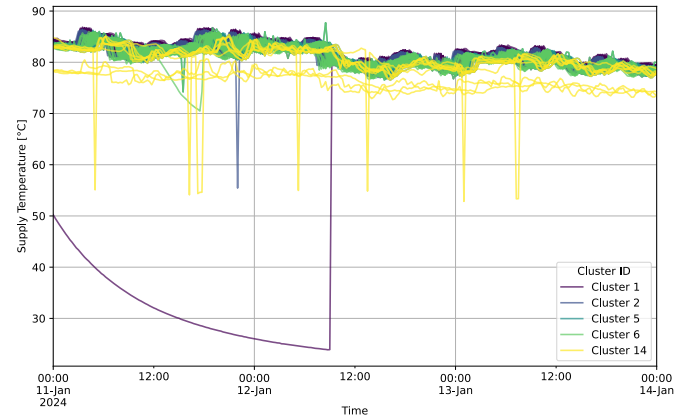

Fig. 4: Clusters of substations. Each line denotes the supply temperature profile of a substation, some of which include noise in the measurement data.

anomalies identified, ten were independently confirmed by a domain expert through manual analysis of the same dataset, underscoring the reliability of our method.

For instance, cluster 24 displays unusual behavior compared to the large share of all substations and is assigned its own cluster. However, this deviation does not necessarily indicate a malfunction. The observed pattern of consistent heat consumption during the daytime with no consumption at night suggests this pattern may indicate the specific operational characteristics or building type associated with this substation. Similarly, Clusters 25 and 27 exhibited unusual behaviors. While consistent heat demand at low supply temperatures is not unusual, Cluster 27 exhibits an anomaly where the return temperature equals the supply temperature, resulting in a ΔT of zero, which is highly unusual and signals further investigation. Other notable instances include Cluster 20, 23, and 30. Cluster 20 shows a higher return temperature than its supply temperature, which may indicate an issue in the system or be related to measurement errors or sensor malfunctions. On the other hand, cluster 23 displays significant fluctuations in supply and return temperatures and may suggest sub-optimal
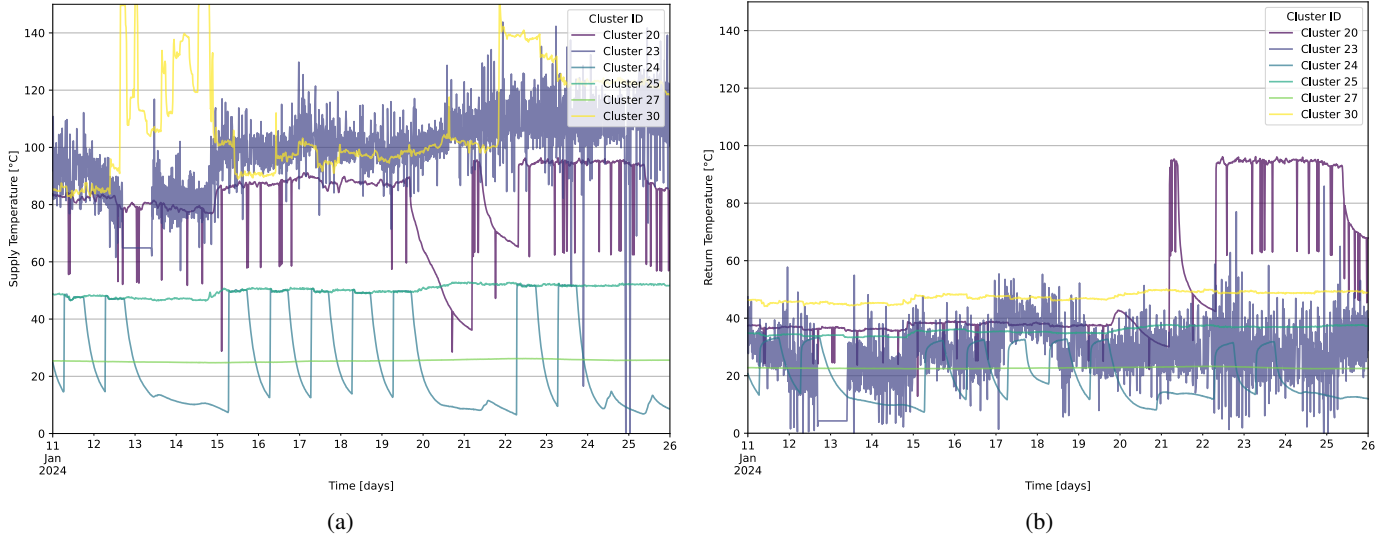
Fig. 5: Subset of supply temperature anomalies. Each substation is shown with its supply temperature (a) and return temperature (b), demonstrating substantial deviations and unusual patterns (cluster 20, 23, 27, and 30), while other substations (cluster 24 and 25) may be unusual though having normal operation.

performance, possibly due to an issue in the substation. Finally, cluster 30 shows a substantial increase in supply temperature, reaching up to 150°C, a condition not observed in other substations. While some substations may show unusual behavior but normal operational patterns, others show signs of more severe conditions. Thus, these supply temperature anomalies could form a basis for strategically focusing maintenance efforts.

B. Performance Anomalies

The relative network approximation forms the basis for our performance anomaly detection. Initially, we identified 30 clusters, from which 16 clusters showing supply temperature anomalies were excluded, leaving us with 14 remaining clusters for the subsequent performance anomaly detection phase. As outlined in subsection IV-B, we construct an MST for each of the 14 remaining clusters. These MSTs enable us to perform intra-cluster performance comparisons. Each substation is evaluated against its k (indirect) neighbors using modified z-scores. In each comparison, the respective substations with a mean modified z-score below 2σ, as compared to their neighboring substations, received an anomaly vote. Votes are normalized by the number of comparisons the substation was involved in, up to $k-1$, resulting in a score between 0 (normal) and 1 (anomalous).

Through this process, we identified 14 substations that received at least one anomaly vote, indicating that compared to their peers, these substations are underperforming. Notably, five of these substations were consistently flagged across all their comparisons, each receiving the highest anomaly score of 1, which indicates a persistently poor performance relative to their peers and across all assessments. Two of these 14 anomalies were independently confirmed by a domain expert through manual analysis of the substations. Ideally, a normal operating substation should show a stable and consistent

pattern in its return temperature or ΔT. Substantial variance or oscillating behavior may hint towards a fault in the substation, ranging from mechanical issues to wrong software settings. Upon further inspection, substations with an anomaly score of 1 exhibited this behavior, including high return temperatures, oscillating readings, low ΔT, or sudden temperature drops, potentially indicating poor heat utilization or other issues such as sensor malfunctions and incorrect settings.

This approach allows for concentrating maintenance efforts on a smaller subset of poorly performing substations. By targeting these specific substations, network operators can enhance overall network performance and reliability and optimize the allocation of maintenance resources. Such an approach enables a cyclical or iterative approach to network maintenance, where utilities can initially focus on the most anomalous substations, e.g., an anomaly score of 1. Post-maintenance evaluations can re-assess the network, focusing on the next set of substations showing significant deviations. This iterative cycle ensures continuous improvement and stability in the operational performance of the DH network while providing an opportunity to collect labeled data on both normal operational behavior and fault instances, e.g., using a labeling taxonomy as outlined in [17]. This collection of labeled data is crucial for advancing the field toward predictive maintenance capabilities instead of reactive fault management.

Our model achieved a sensitivity of approximately 65% and specificity of approximately 97%. This score suggests potential areas for improvement in identifying anomalies but demonstrates a robust capacity to correctly identify non-anomalous substations and minimize false positives. This high level of specificity is critical as it helps to avoid unnecessary maintenance expeditions, thereby conserving resources and optimizing operational efficiency.

C. Limitations

While our methodology has demonstrated effectiveness in network topology approximation and anomaly detection of substations, some limitations should be considered. The performance of our method is sensitive to the choice of hyperparameters, such as the number of neighbors, warp distance, and edge inclusion/exclusion thresholds, which require specific domain knowledge. Moreover, while the method's performance might vary across different DH networks, it is based on universal aspects of DH systems. We expect the SHEDAD approach to be generalizable to other DH networks.

VI. Conclusions and Future Work

This study introduced a novel SNN-enhanced approach (SHEDAD) for anomaly detection in DH networks, leveraging the relative network topology to identify operational anomalies. Our method adeptly manages the complex data structures of DH systems and maintains network confidentiality. It has effectively isolated anomalies, identifying 16 supply temperature anomalies and 14 performance anomalies among 248 substations. We achieved a sensitivity of approximately 65% and a specificity of approximately 97%. The results underscore the potential of our automatic method to improve the efficiency and reliability of DH networks by enabling more precise maintenance responses based on the identified anomalies. The proposed approach promotes a cyclical strategy for network maintenance, prioritizing intervention at substations with the highest anomaly scores. This iterative process enhances the operational performance of DH networks and facilitates the systematic collection of labeled data of both normal and faulty data. Such data are invaluable for training and validating data-driven models, improving anomaly detection and fault diagnosis capabilities over time. Possible future efforts are directed towards the following:

- Enhancing the method's network approximation capabilities, aiming to develop uniform-sized clusters and dynamically adjust the number k more effectively.
- Extending validation and testing across various types of DH networks to improve adaptability and effectiveness under diverse conditions. We plan to integrate federated learning to enable decentralized, collaborative anomaly detection across substations without compromising data privacy.
- Enhancing anomaly detection processes via machine learning techniques that harness federated learning for model training on data collected through targeted maintenance operations.

References

[1] Ece Calikus et al. Ranking abnormal substations by power signature dispersion. *Energy Procedia*, 149:345–353, Sep 2018.

[2] Ece Calikus et al. A data-driven approach for discovering heat load patterns in district heating. *Applied Energy*, 252:113409, Oct 2019.

[3] Teri Crosby. How to detect and handle outliers. *Technometrics*, Aug 1994.

[4] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[5] Henrik Gadd and Sven Werner. Fault detection in district heating substations. *Applied Energy*, 157:51–59, Nov 2015.

[6] Panagiota Gianniou et al. Clustering-based analysis for residential district heating data. *Energy Conversion and Management*, 165:840–850, Jun 2018.

[7] Yejin Hong and Sungmin Yoon. Holistic operational signatures for an energy-efficient district heating substation in buildings. *Energy*, 250:123798, Jul 2022.

[8] IEA. 350 million building units connected to district energy networks by 2030, provide about 20% of space heating needs. URL Link, 2022. Accessed: 2023-08-01.

[9] Paul Jaccard. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50, 1912.

[10] R.A. Jarvis and E.A. Patrick. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers*, C–22(11):1025–1034, Nov 1973.

[11] Minseok Kim, Seunghwan Jung, Eunkyeong Kim, Baekcheon Kim, Jinyong Kim, and Sungshin Kim. A fault detection and isolation method via shared nearest neighbor for circulating fluidized bed boiler. *Processes*, 11:3433, December 2023.

[12] Jad Al Koussa and Sara Månsson. Fault detection in district heating substations: a cluster-based and an instance-based approach. Apr 2022.

[13] Daniel Leiria et al. Towards automated fault detection and diagnosis in district heating customers: generation and analysis of a labeled dataset with ground truth. volume 18 of *Building Simulation*, page 3620–3628. IBPSA, 2023.

[14] Rui Liu et al. Shared-nearest-neighbor-based clustering by fast search and find of density peaks. *Information Sciences*, 450:200–226, Jun 2018.

[15] Sara Månsson et al. A machine learning approach to fault detection in district heating substations. *Energy Procedia*, 149:226–235, Sep 2018.

[16] Sara Månsson et al. Automated statistical methods for fault detection in district heating customer installations. *Energies*, 12(11):113, Jan 2019.

[17] Sara Månsson et al. A taxonomy for labeling deviations in district heating customer data. *Smart Energy*, 2:100020, May 2021.

[18] Martin Neumayer et al. Fault and anomaly detection in district heating substations: A survey on methodology and data sets. *Energy*, 276:127569, Aug 2023.

[19] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, Feb 1978.

[20] Souhardya Sengupta and Swagatam Das. Selective nearest neighbors clustering. *Pattern Recognition Letters*, 155:178–185, Mar 2022.

[21] RR Sokal and CD Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin. University of Kansas*, 1958.

[22] Felix T. et al. Fault detection and condition monitoring in district heating using smart meter data. *PHM Society European Conference*, 6(1):11, Jun 2021.

[23] Alexander Martin Tureczek et al. Clustering district heat exchange stations using smart meter consumption data. *Energy and Buildings*, 182:144–158, Jan 2019.

[24] Jonne van Dreven et al. Intelligent approaches to fault detection and diagnosis in district heating: Current trends, challenges, and opportunities. *Electronics*, 12(66):1448, Jan 2023.

[25] Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, Mar 1963.

[26] Sven Werner. International review of district heating and cooling. *Energy*, 137:617–631, Oct 2017.

[27] Chunrong Wu et al. Effective hierarchical clustering based on structural similarities in nearest neighbor graphs. *Knowledge-Based Systems*, 228:107295, Sep 2021.

[28] Puning Xue et al. Fault detection and operation optimization in district heating substations based on data mining techniques. *Applied Energy*, 205:926–940, 2017.

[29] Jing Zhao and Yu Shan. An influencing parameters analysis of district heating network time delays based on the cfd method. *Energies*, 12(77):1297, Jan 2019.