Sensor Fault Detection via Virtual Smart Heat Metering with Spatial-Temporal Graph Neural Networks

Keivan Faghih Niresi[1], and Olga Fink[2]

[1,2] *EPFL, Switzerland*
keivan.faghihniresi@epfl.ch
olga.fink@epfl.ch

ABSTRACT

Sensor faults and miscalibrated sensors represent an important challenge in district heating networks, where measurement errors, drift, calibration inaccuracies, or communication issues can compromise the reliability of thermal and hydraulic monitoring. Detecting such issues in a timely manner is essential for maintaining operational efficiency and ensuring the trustworthiness of system data. A promising approach for addressing these challenges is to compare physical measurements with estimates generated by virtual sensors. Virtual sensing enables the reconstruction of unmeasured or unreliable variables using data-driven models and existing measurements, thereby providing an estimate of the expected measurement value under the current operating and environmental conditions, which can serve as reference against which anomalous or inconsistent sensor behavior can be identified. In this work, we develop a virtual-sensor–based framework for sensor fault and miscalibration detection using a heterogeneous spatial–temporal graph neural network (HSTGNN). The proposed model learns both the spatial relationships among sensors and the temporal dynamics of their measurements to construct accurate virtual smart heat meter outputs. To evaluate the approach, we use a controlled laboratory dataset collected at the Aalborg Smart Water Infrastructure Laboratory, which provides synchronized high-resolution measurements of flow, temperature, and pressure representative of district heating operating conditions. Experimental results demonstrate that the proposed HSTGNN improves fault detection performance compared to several baseline methods.

1. INTRODUCTION

District heating systems can significantly improve urban energy efficiency by facilitating the large-scale integration of renewable and waste heat sources (Huang et al., 2020). The reliable operation and monitoring of these systems rely heavily on sensor measurements, including flow rate, temperature, and pressure, which provide essential information about the thermal and hydraulic state of the network (Niresi, Bissig, Baumann, & Fink, 2024). However, sensors deployed in district heating infrastructures are often subjected to harsh operating conditions and long service lifetimes. Over time, these factors can lead to various measurement issues, such as drift, bias, spikes, or communication interruptions (Belgacem & Chihi, 2025). Such inaccuracies can compromise the quality of monitoring data and may ultimately lead to suboptimal or incorrect operational decisions. Consequently, the timely detection of faulty or unreliable sensor measurements is crucial for maintaining reliable and trustworthy system operation.

A common strategy for detecting faulty measurements is to compare sensor readings with expected system behavior derived from physical models or data-driven estimators. If the observed measurement deviates significantly from the expected value under the current operating conditions, the sensor may be malfunctioning or miscalibrated (Hsu, Frusque, & Fink, 2023). However, developing accurate reference models for complex infrastructure systems such as district heating networks is challenging due to their spatially distributed structure and dynamic operating conditions. One practical way to obtain such reference values is to estimate the expected measurement directly from system data by exploiting relationships between variables in the network. Virtual sensing represents an effective approach for this purpose. A virtual sensor estimates the value of a physical variable by leveraging available measurements and learned relationships within the system, thereby improving system observability without requiring additional hardware installations (Sun & Ge, 2021). When a virtual sensor provides an accurate estimate of the expected measurement under current operating conditions, deviations between the virtual and physical sensor outputs can be used to identify abnormal behavior and detect faulty sensors. In this way, virtual sensing provides a practical and scalable approach for sensor fault detection in complex infrastructure systems such as district heating networks (Fink, Nejjar, et al., 2026).

The effectiveness of sensor fault detection strongly depends on the fidelity of the virtual estimate. When a virtual sensor closely represents the true reference under current operating conditions, deviations between the virtual and physical sensors can more reliably indicate sensor faults. For distributed sensor networks such as district heating systems, architectures that encode the network's spatial structure provide a useful inductive bias because they exploit physical or functional connectivity between sensors (Fink, Sharma, et al., 2026); graph neural networks (GNN) are a natural way to impose that bias. Standard GNNs excel at learning spatial relationships in graph-structured data (Wu et al., 2020). While standard GNNs are effective at capturing spatial dependencies in such graph-structured systems, measurements in district heating networks also evolve over time and exhibit strong temporal dynamics. To account for both spatial interactions and temporal evolution, spatial-temporal graph neural networks (STGNNs) have been proposed. These models are particularly well-suited for this setting, as they can simultaneously capture spatial dependencies between sensors and temporal dynamics in measurements (Jin et al., 2024). STGNN-based models can therefore generate highly accurate predictions of system variables (Theiler & Fink, 2025) and enable reliable virtual sensing (Niresi, Nejjar, & Fink, 2025) by learning these spatial-temporal patterns.

Standard STGNN architectures generally do not explicitly account for node heterogeneity (Zhao, Taal, Baggerohr, & Fink, 2025) and often assume homogeneous node features and shared model parameters across all nodes in the graph. However, district heating systems contain heterogeneous sensor types, such as flow, temperature, and pressure sensors, each exhibiting distinct physical behaviors, spatial dependencies, and temporal dynamics. Ignoring these differences may limit the model's ability to accurately capture the underlying system dynamics and detect abnormal measurements.

In this paper, we propose a heterogeneous spatial–temporal graph neural network (HSTGNN) framework for virtual sensing-enabled fault detection in district heating systems. The proposed architecture explicitly accounts for the heterogeneous characteristics of different sensor types, including flow, temperature, and pressure measurements. Specifically, the model consists of three sensor-specific streams that learn distinct spatial relationships and temporal dynamics for each sensor type. The learned representations are subsequently fused through a fusion module to estimate virtual smart heat meter outputs. Sensor faults are then detected by analyzing deviations between the virtual sensor estimates and the corresponding physical measurements.

To evaluate the proposed approach, experiments are conducted using a controlled laboratory dataset collected at the *Aalborg Smart Water Infrastructure Laboratory* (Val Ledesma, Wisniewski, & Kallesøe, 2021), which provides synchronized high-resolution measurements of flow, temperature, and pressure representative of district heating operating conditions. Experimental results demonstrate that the proposed HSTGNN significantly improves virtual fault detection accuracy compared to other methods.

The main contributions of this work are summarized as follows:

- We propose a virtual-sensor–based framework for sensor fault detection in district heating systems.
- We develop a sensor-type–aware spatial–temporal graph neural network that captures heterogeneous spatial and temporal relationships among flow, temperature, and pressure sensors.
- We demonstrate how deviations between virtual and physical sensor measurements can be exploited to detect sensor faults without requiring additional hardware.
- We validate the proposed method using a high-resolution experimental dataset representative of district heating operating conditions.

The results highlight the potential of graph-based virtual sensing methods to improve monitoring reliability and support detection of faulty sensors in district heating networks.

2. RELATED WORKS

Virtual sensing methods for district heating are commonly grouped into three categories: physics-based simulation, purely data-driven models, and hybrid approaches that combine simple physical relationships with learned components. Physics-based estimates can provide physically consistent and interpretable estimates that are valuable for monitoring and fault detection of sensors, but building and maintaining high-fidelity digital twins (i.e., virtual replicas of physical assets) at scale is costly, labor-intensive, and may not always be possible (Bank, Madsen, Mortensen, Søndergaard, & Shaker, 2023). When such models are unavailable or impractical, data-driven approaches offer an alternative. However, many implemented models are conventional black-box architectures such as multilayer perceptrons (MLPs) that do not explicitly leverage the network's spatial topology or the temporal dynamics of sensor signals (Darvishi, Ciuonzo, Eide, & Rossi, 2020). The current predominant hybrid approaches combine simple physical relationships with learned components, providing a balance between interpretability and flexibility. They can improve robustness under varying operating conditions by incorporating domain knowledge (Yoon et al., 2020). However, both the physical priors and the data-driven components of the previous approaches are relatively simple and do not fully capture the complex dynamics of district heating networks in hybrid models. As a result, these models often fail to exploit key inductive, learning, and observational biases (Fink, Sharma, et al., 2026; Fink, Nejjar, et al., 2026).

Recent efforts have therefore turned to graph-based learning to exploit the natural network structure of district heating systems and to capture spatial–temporal signal propagation. Approaches that combine graph inference (Niresi, Kuhn, Frusque, & Fink, 2024) or physics-derived graph augmentations with graph neural networks (Niresi, Bissig, et al., 2024) have shown promise for soft sensing by enforcing spatial consistency and modeling temporal evolution. However, existing graph-based methods often share two practical limitations: evaluations are frequently performed on simulated or synthetic datasets rather than extensive real-world measurements, and measured signals in district heating networks are inherently heterogeneous. Few approaches have accounted for this complexity (Zhao et al., 2025), as many standard spatial–temporal GNN formulations assume homogeneous node features and shared parameters, thereby reducing their ability to represent node-level heterogeneity such as varying sensor modalities.

3. METHODOLOGY

3.1. Sensor-Type-Specific Modeling

Virtual sensing in district heating must account for the fact that different sensor types measure distinct physical quantities that follow different dynamics, noise characteristics, and spatial-temporal relationships. Moreover, operating and environmental conditions, such as network load or ambient temperature, can affect these measurements differently depending on the underlying physical process. Temperature sensors primarily reflect thermal inertia and slow transients, flow sensors respond to hydraulic changes and often display faster dynamics, and pressure sensors capture compressive and network-wide propagation effects. These differences imply that a single, homogeneous processing pipeline will tend to underfit type-specific behaviour, which reduces the fidelity of the virtual estimates and degrades downstream fault detection.

For this reason, we treat temperature, flow, and pressure sensors separately in the model design. Separate temporal encoders allow the network to learn type-specific response times, parameter sharing within each type improves sample efficiency by pooling information among sensors with similar physics, and dedicated intra-type graph learning captures the most relevant local interactions before integrating information across types. Cross-type integration is still necessary because physical couplings exist (for example, changes in flow can lead to temperature variations), but modelling these couplings after type-specific processing makes the fused representation more physically plausible and easier to interpret.

3.2. Problem Formulation

Let $\mathbf{X} \in \mathbb{R}^{N \times T}$ denote the multivariate time-series input over a window of length T for N sensor nodes. The sensor network is represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where nodes $v_i \in \mathcal{V}$ represent individual physical sensors and $\mathbf{A} \in \mathbb{R}^{N \times N}$ denotes the adjacency matrix encoding the strength of connections between sensors. Each node belongs to one of three physical sensor types: temperature, pressure, or flow. Hence, the node set $\mathcal{V}$ is partitioned into three disjoint subsets according to the physical sensor type: temperature ($\mathcal{V}_\mathrm{T}$), pressure ($\mathcal{V}_\mathrm{P}$), and flow ($\mathcal{V}_\mathrm{F}$):

$$\mathcal{V} = \mathcal{V}_\mathrm{T} \cup \mathcal{V}_\mathrm{P} \cup \mathcal{V}_\mathrm{F}, \qquad \mathcal{V}_i \cap \mathcal{V}_j = \emptyset \quad \forall i \neq j.$$

Accordingly, the total number of sensors satisfies $N = N_\mathrm{T} + N_\mathrm{P} + N_\mathrm{F}$. Edges in $\mathcal{E}$ represent dependencies between sensors, capturing correlations and potential physical interactions within the network. These relationships may reflect both direct physical couplings and latent influences learned from data.

The objective of the virtual sensing model in this work is to generate reliable reference estimates for selected sensors in the network, which are subsequently used for residual-based fault detection. The model predicts the expected measurements of target sensors based on historical observations from the remaining sensors in the network. These predictions represent the expected normal behaviour of the system under the current operating conditions. Potential sensor faults can then be identified by comparing the predicted virtual sensor values with the corresponding physical measurements and analysing the resulting residuals.

The residual between the measured and predicted sensor value is defined as

$$r_i(t) = y_i(t) - \hat{y}_i(t),$$

where $y_i(t)$ denotes the measured value of sensor i at time t, and $\hat{y}_i(t)$ denotes the corresponding virtual sensor estimate produced by the model.

Formally, the model learns an estimator that takes a window of historical measurements from temperature, pressure, and flow sensors and predicts the expected measurements of the target smart meters at the current time step:

$$\hat{\mathbf{y}}_t = f\left(\mathbf{X}_\mathrm{T}^{t-T:t}, \mathbf{X}_\mathrm{P}^{t-T:t}, \mathbf{X}_\mathrm{F}^{t-T:t}\right) \tag{1}$$

where $\hat{\mathbf{y}}_t$ denotes the virtual sensor estimates produced by the model. These estimates serve as reference signals for the residual-based fault detection procedure described later.

3.3. Proposed HSTGNN Architecture

Building on our prior HSTGNN virtual sensor architecture (Niresi, Jensen, Kallesøe, Wisniewski, & Fink, 2026), we extend it with a residual-based fault detection scheme. The model uses dedicated branches per sensor type to learn type-specific temporal dynamics and intra-type spatial relationships. The representations from the different branches are subsequently integrated through an attention-based fusion mechanism to capture cross-type interactions. A branch corresponds to the complete processing pathway for a single sensor type and includes a temporal encoder with type-shared parame-

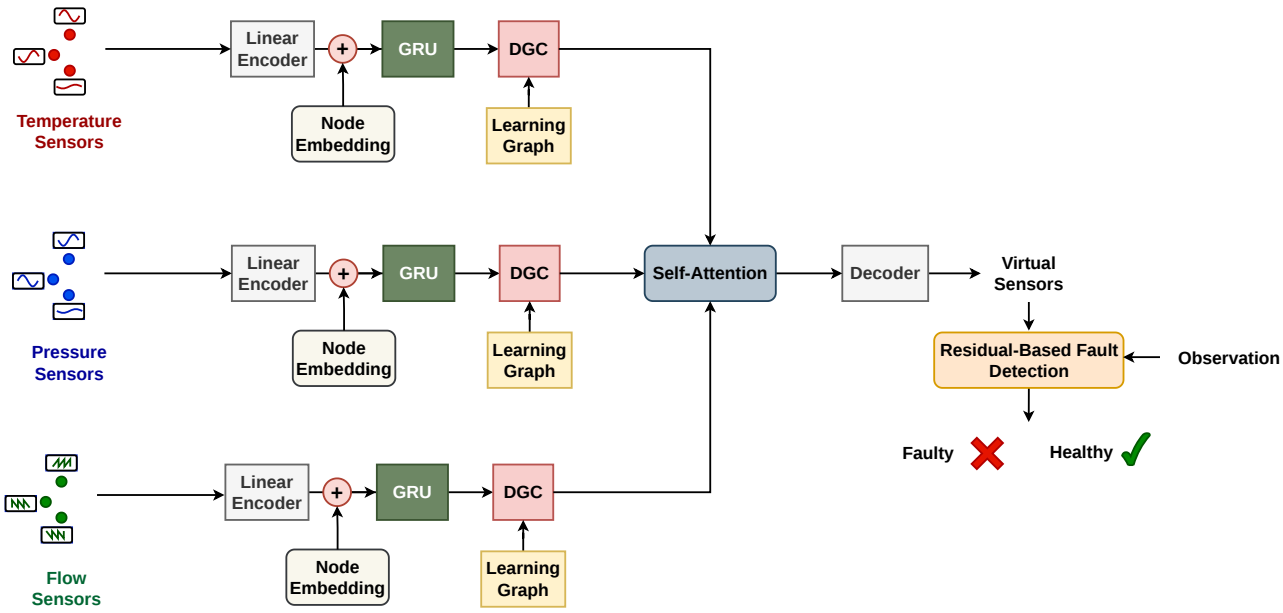

Figure 1. The Proposed HSTGNN Architecture for Sensor Fault Detection

ters, latent intra-type graph learning, and spatial aggregation. By separating the processing of temperature, flow, and pressure sensors, the model can capture their distinct dynamics and measurement characteristics before modeling cross-type couplings. A schematic overview of the proposed HSTGNN framework, highlighting the sensor-type–specific processing streams, and the sensor fault detection workflow, is shown in Fig. 1.

Temporal Modeling and Sensor-Specific Embedding The initial processing stage aims to capture two key aspects: the distinct temporal dynamics of different physical variables and the node-specific characteristics of each sensor. Even sensors of the same physical type may exhibit different dynamics depending on their location within the district heating network. To address this, the raw input measurements are first projected into a hidden feature space through a linear transformation. At this stage, sensors may exhibit similar historical measurement patterns. To allow the model to distinguish between sensors and incorporate contextual information about their position in the network, learnable node embeddings are added to the projected features. Temporal dependencies are then modeled independently for each node v_i using Gated Recurrent Units (GRUs):

$$\mathbf{h}_i^t = \mathrm{GRU}(x_i^{t-T:t}; \Theta_{\tau(v_i)}) \tag{2}$$

where $\Theta_{\tau(v_i)}$ represents parameters shared among sensors of the same physical type, allowing the model to capture type-specific temporal characteristics. This enables the model to learn dynamics such as the slower thermal inertia typically observed in temperature measurements or the faster response dynamics associated with hydraulic variables like flow and pressure.

Latent Graph Learning Rather than relying solely on the physical network topology, which may not fully capture the effective interactions between sensors, the intra-type spatial structure is learned directly from data. In district heating systems, the influence between sensors is not determined only by the physical pipe connectivity but also by hydraulic dynamics, control actions, and varying operating conditions. As a result, sensors that are physically distant may exhibit strong correlations, while physically connected sensors may interact weakly depending on the system state.

To capture these effective dependencies, we learn a score matrix $\Phi^{(m)}$ for each sensor type m, representing potential interactions between sensors of the same type (Cini, Zambon, & Alippi, 2023). For each node i, a subset of K neighbors is selected using the Gumbel-TopK trick (Kool, Van Hoof, & Welling, 2020):

$$M_i \sim \mathrm{Categorical}\left(\mathrm{Softmax}(\Phi_i^{(m)})\right) \tag{3}$$

This procedure enables discrete neighbor selection during the forward pass while preserving differentiability through a straight-through gradient estimator (Bengio, Léonard, & Courville, 2013). As a result, the model can learn sparse and informative spatial dependencies that reflect the effective information flow within the network.

Spatial Aggregation via Diffusion Spatial dependencies are modeled using Diffusion Graph Convolution (DGC), which propagates information across the learned graph structure. The

diffusion process is defined as (Atwood & Towsley, 2016; Li, Yu, Shahabi, & Liu, 2017):

$$\mathbf{H} = \sum_{s=0}^{S} \psi \left((\mathbf{D}^{-1}\mathbf{A})^s \mathbf{X}\mathbf{W}^{(s)} \right) \tag{4}$$

where $\mathbf{D}^{-1}\mathbf{A}$ denotes the state transition matrix, $\mathbf{W}^{(s)}$ are learnable diffusion weights, and S represents the number of diffusion steps. This formulation enables information to propagate across multiple hops in the graph, allowing the model to capture spatial dependencies beyond immediate neighbors. To account for the bidirectional nature of hydraulic influences in district heating networks, diffusion is applied to both the transition matrix $\mathbf{D}^{-1}\mathbf{A}$ and its transpose. This allows the model to capture dependencies corresponding to both upstream and downstream propagation of hydraulic and thermal effects.

Cross-Type Attention and Decoding After the intra-type processing stages, the representations of all sensor nodes are combined and processed using a self-attention mechanism (Vaswani et al., 2017). This layer enables the model to capture cross-type interactions between different physical variables, such as the influence of flow rate variations on temperature dynamics. The resulting node representations are then aggregated and passed through a linear decoder to produce the virtual sensor estimates $\hat{\mathbf{y}}_t$.

Training Objective The model is trained by minimizing the Mean Absolute Error (MAE) between the predicted and observed sensor values:

$$\mathcal{L} = \frac{1}{D} \sum_{j=1}^{D} |y_{t,j} - \hat{y}_{t,j}| \tag{5}$$

where $y_{t,j}$ and $\hat{y}_{t,j}$ represent the ground-truth and predicted values for the j-th target sensor at time step t, respectively. The MAE loss is chosen due to its reduced sensitivity to large errors compared to quadratic losses.

3.4. Fault Detection

To enable detection of both abrupt (bias) and gradual (drift) deviations in sensor networks, we adopt a residual-based detection scheme. For each sensor s at time step t, the residual $r_{t,s}$ is defined as the absolute difference between the observed measurement and the corresponding predicted value obtained from a reference model (e.g., a virtual sensor):

$$r_{t,s} = \left| y_{t,s}^{\text{obs}} - y_{t,s}^{\text{vs}} \right|, \tag{6}$$

where $y_{t,s}^{\text{obs}}$ denotes the observed sensor reading and $y_{t,s}^{\text{vs}}$ the virtual sensor estimate. The residual thus quantifies the instantaneous deviation indicative of a potential fault.

A deviation is flagged whenever the residual exceeds a sensor-specific threshold τ_s:

$$\hat{y}_{t,s}^{\text{dev}} = \begin{cases} 1, & r_{t,s} > \tau_s \\ 0, & \text{otherwise} \end{cases}. \tag{7}$$

The selection of the threshold τ_s is critical for balancing detection sensitivity against the false alarm rate. In this work, we determine the optimal threshold for each sensor individually using a validation dataset. Specifically, we evaluate a range of potential threshold values and select the τ_s that maximizes the F_1 score based on the validation data:

$$\tau_s = \arg\max_{\tau} F1(\tau; \mathcal{D}_{\text{val},s}), \tag{8}$$

where $\mathcal{D}_{\text{val},s}$ represents the validation residuals for sensor s. By maximizing the $F1$ score (the harmonic mean of precision and recall), we ensure that the detection scheme effectively identifies true faults while minimizing spurious detections caused by inherent modeling noise. This detection procedure is applied independently to each scenario considered: bias, drift, and mixed, and across all baseline models used for comparison.

3.5. Baseline Methods

To evaluate the effectiveness of the proposed HSTGNN model for sensor fault detection, we compare it with four established baselines commonly used for time-series modelling. The baselines include a Long Short-Term Memory (LSTM) network (Hochreiter & Schmidhuber, 1997), a one-dimensional convolutional neural network (1D-CNN), Diffusion Graph Convolution (DGC) (Atwood & Towsley, 2016; Li et al., 2017), and a Graph Convolutional Network (GCN) (Kipf & Welling, 2016). The LSTM model captures temporal dependencies in sequential data and has been widely applied in industrial monitoring tasks. The 1D-CNN relies on convolutional filters to extract local temporal patterns and has demonstrated strong performance in short-term anomaly detection. DGC incorporates relational dependencies between sensors through diffusion-based graph convolutions, enabling the model to exploit spatial correlations within the network. Similarly, GCN models inter-sensor relationships using graph convolutional layers and have been shown effective in learning structural dependencies in sensor networks. These baselines represent temporal-only, convolutional, and graph-based approaches, providing a diverse set of comparison models for evaluating the proposed HSTGNN, which jointly captures temporal dynamics and spatial dependencies.

4. CASE STUDY

To evaluate the proposed virtual sensing and fault detection framework, we consider a set of sensor fault scenarios repre-

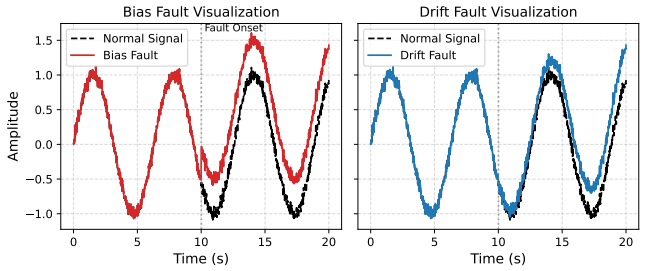

Figure 2. Comparison of simulated sensor fault models against a normal signal.

sentative of common measurement errors in district heating networks. In particular, we investigate three types of deviations: bias, drift, and mixed faults. A bias fault corresponds to a sudden offset added to the sensor measurement, representing abrupt calibration errors or sensor malfunctions. A drift fault represents a gradual deviation that increases over time, reflecting progressive sensor degradation. The mixed scenario combines both effects, where a gradual drift is followed by an abrupt shift. To illustrate the considered fault types, Fig. 2 provides a schematic example of bias and drift faults.

Faults are injected into selected sensors to simulate deviations from normal operation, while the remaining sensors provide contextual information for the virtual sensing models. The models are trained on data representing normal operating conditions and evaluated on datasets containing the injected faults. In the Bias scenario, a constant offset corresponding to 10% of each sensor's measurement range was added to the signal. The Drift scenario introduced gradually increasing deviations with a slope magnitude corresponding to 20% of the sensor's range, simulating slowly developing sensor miscalibration. The Mixed scenario combines both fault types, where bias faults are injected into the first three sensors and drift faults into the remaining three, creating a heterogeneous fault environment representative of a challenging scenario in sensor networks. This setup allows us to assess the ability of the proposed method to detect both sudden and slowly evolving sensor faults under realistic operating conditions.

For each scenario, residual-based fault detection is applied by comparing the measured sensor values with the corresponding virtual sensor estimates. Detection performance is evaluated across all models and scenarios using the residual thresholding procedure described in the previous section.

4.1. Testbed and Data Acquisition

Experimental data were collected at the Aalborg University Smart Water Infrastructures Laboratory (SWIL), which implements a tree-structured district heating network topology. The network contains two consumer stations, each equipped with a *Kamstrup Multical 303* smart meter (Kamstrup, 2026). Each smart meter provides three measurements: the flow rate, inlet temperature, and outlet temperature at the consumer substation. Consequently, the smart meters provide a total of six reference output variables corresponding to the two consumers. In addition to the smart meters, which are treated as high-accuracy reference measurements, the rest of the network is instrumented with pressure, flow, and temperature sensors distributed throughout the network, following the layout illustrated in (Niresi et al., 2026). These sensors provide input measurements for the learning models and may be affected by measurement errors such as miscalibration or sensor drift. Data acquisition is performed using Beckhoff I/O modules integrated with MATLAB/Simulink, with a Real-Time Pacer ensuring a uniform sampling rate of $f_s = 0.5$ Hz. The dataset consists of approximately eight hours of continuous measurements, capturing the dynamic behaviour of the network under normal operating conditions. During this period, the network was supplied by a central boiler and pump station, which maintained a supply temperature of approximately 45°C. The return temperature was observed at around 35°C.

The dataset was split chronologically into 70% training, 10% validation, and 20% test sets. To prevent data leakage, the mean μ_{train} and standard deviation σ_{train} are computed exclusively from the training set for Z-score standardization:

$$\hat{x} = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}}. \tag{9}$$

5. EXPERIMENTAL RESULTS

5.1. Performance Metrics

To comprehensively evaluate the effectiveness of the proposed HSTGNN model and the baseline methods, five standard classification metrics were employed: Accuracy, Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC). These metrics are calculated based on the outcomes of the confusion matrix: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). In the context of our fault detection framework, a positive instance indicates a faulty sensor reading, while a negative instance represents normal system behavior.

The metrics are defined as follows:

- **Accuracy:** The proportion of correctly classified instances (both normal and faulty) out of the total dataset.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- **Precision:** The ratio of correctly predicted faulty readings to all instances predicted as faulty by the model. A high precision score indicates a low false alarm rate, meaning that when the system raises an alert, it is highly likely to

be a genuine fault.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity):** The proportion of actual faulty readings that were successfully identified by the model. High recall is critical in safety-critical systems to ensure that incipient or hidden faults do not go unnoticed.

$$Recall = \frac{TP}{TP + FN}$$

- **F1-score:** The harmonic mean of Precision and Recall. It provides a single, unified metric that balances the trade-off between false positives and false negatives. It is an especially robust evaluation metric when dealing with imbalanced data where standard accuracy falls short.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- **AUC (Area Under the Curve):** The AUC measures the entire two-dimensional area underneath the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the True Positive Rate (Recall) against the False Positive Rate ($FPR = \frac{FP}{FP+TN}$) across various classification thresholds. An AUC closer to 1.0 signifies a superior measure of separability, indicating the model's robust ability to distinguish between faulty and normal states regardless of the chosen threshold.

5.2. Hyperparameter and Model Configuration

Hyperparameters and configurations were identified via a systematic grid search on the validation dataset to ensure experimental integrity and prevent data leakage across subsequent tests. These parameters are summarized in Table 1. Temporal dependencies were captured using a sliding window of length $T = 16$ with a stride of 1, providing an optimal balance between predictive fidelity and computational efficiency. The final architecture utilizes a 2-layer GRU with 16 hidden units, followed by a single-layer GNN ($K = 8$ neighbors) with 16 output channels. Training was conducted using the Adam optimizer with a learning rate of 0.001 and a batch size of 512, minimizing the MAE to ensure robust convergence.

To maintain a fair comparison, all baseline models follow the same procedure. The LSTM baseline uses two stacked layers with a 256-dimensional hidden state, while the 1D-CNN utilizes 64 filters followed by adaptive pooling and a linear output mapping. For graph-based methods, we fix the latent node representation at a size of 16, combining projected node features with learnable embeddings of the same dimensionality. Specifically, the DGC model employs a two-step diffusion convolution, and the GCN uses a standard graph convolution, both feeding into a linear decoder.

Table 1. Configuration details: Exploratory search space and final model hyperparameters.

Modules	Hyperparameter	Search Space (Optimal in Bold)
Input Processing	Window Size	{4, 8, **16**, 32, 64}
	Stride	1 (Fixed)
Graph	GNN Layers	{**1**, 2, 3, 4}
	Output Channels	{4, 8, **16**, 32, 64}
	Kernel Size	{1, **2**, 3, 4}
	Node Embedding	{4, 8, **16**, 32, 64}
	K neighbors	{2, 4, 6, **8**, 10, 12}
Recurrent	GRU Layers	{1, **2**, 3, 4}
	Hidden Units	{4, 8, **16**, 32, 64}
Training	Batch Size	512
	Loss Function	Mean Absolute Error (MAE)
	Optimizer	Adam
	Learning Rate	0.001

Table 2. Performance Metrics under Bias Fault Scenario

Method	Precision	Recall	F1	Accuracy	AUC
LSTM	0.923	0.955	0.939	0.937	0.956
1D-CNN	0.911	0.926	0.918	0.918	0.943
DGC	0.769	0.932	0.842	0.826	0.882
GCN	0.786	0.928	0.851	0.838	0.886
HSTGNN	**0.956**	**0.970**	**0.963**	**0.962**	**0.989**

5.3. Sensor Fault Detection Performance

In the Bias scenario (Table 2), which simulates abrupt deviations from normal sensor behavior, most models achieve relatively high detection performance. The proposed HSTGNN model achieves the highest overall performance with an F1-score of 0.963 and an AUC of 0.989, indicating highly accurate identification of biased readings with minimal false positives. Among the baselines, the LSTM performs strongly with F1-score of 0.939, suggesting that abrupt faults can be effectively detected when temporal dependencies are modeled. The 1D-CNN also performs reasonably well, achieving an F1-score of 0.918, while GCN and DGC achieve a lower F1-score of 0.851 and 0.842, respectively, highlighting their limitations in capturing temporal characteristics of the sensor signals.

The Drift scenario (Table 3) presents a more challenging detection task. Drift faults, characterized by gradual changes, are easily obscured by natural temporal variations in the data. Consequently, all models experienced a performance drop compared to the Bias scenario. HSTGNN demonstrated superior robustness, maintaining an F1-score of 0.854 and an AUC of 0.907. Baseline models struggled significantly with both Precision and Recall, with F1-scores ranging from 0.768 (DGC) to 0.811 (LSTM). The low Recall of many baselines indicates a high rate of false negatives, meaning subtle drift faults were frequently missed. HSTGNN's Recall of 0.839 confirms that its integrated spatial-temporal modeling effectively distinguishes gradual anomalies from normal system dynamics.

Table 4 details the performance under the Mixed fault scenario,

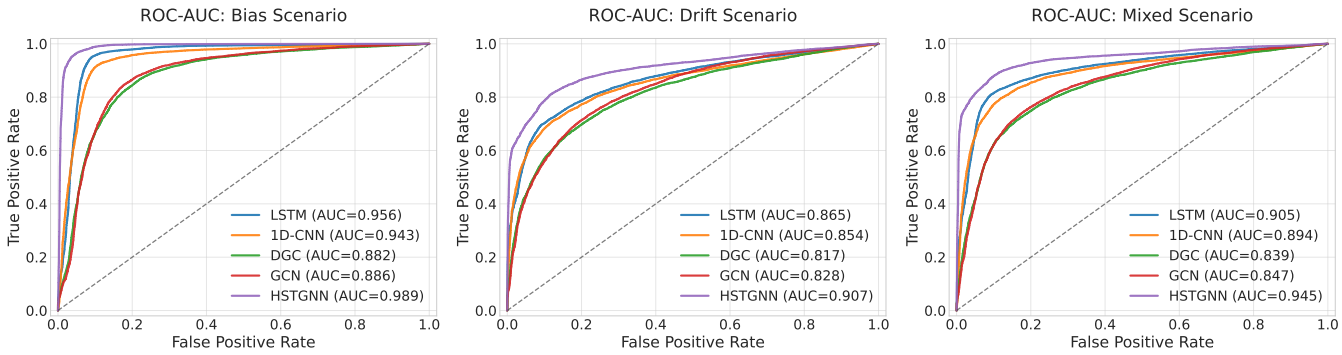

Figure 3. ROC curves for all models under the Bias, Drift, and Mixed fault scenarios. HSTGNN consistently achieves the highest True Positive Rate across nearly all False Positive Rate levels, demonstrating superior discriminative ability in detecting both abrupt and gradual faults compared to baseline models.

Table 3. Performance Metrics under Drift Fault Scenario

Method	Precision	Recall	F1	Accuracy	AUC
LSTM	0.831	0.793	0.811	0.816	0.865
1D-CNN	0.822	0.783	0.802	0.807	0.854
DGC	0.740	0.798	0.768	0.759	0.817
GCN	0.748	0.820	0.782	0.772	0.828
HSTGNN	**0.869**	**0.839**	**0.854**	**0.856**	**0.907**

Table 4. Performance Metrics under Mixed Fault Scenario

Method	Precision	Recall	F1	Accuracy	AUC
LSTM	0.880	0.860	0.870	0.871	0.905
1D-CNN	0.860	0.848	0.854	0.855	0.894
DGC	0.751	0.848	0.797	0.784	0.839
GCN	0.764	0.854	0.806	0.795	0.847
HSTGNN	**0.930**	**0.904**	**0.917**	**0.918**	**0.945**

representing a highly realistic environment with both abrupt and gradual faults occurring simultaneously. HSTGNN again outperformed all baselines, achieving a Precision of 0.930, a Recall of 0.904, and an overall F1-score of 0.917. Its ability to maintain high precision while capturing the majority of faults highlights its capacity to disentangle overlapping anomaly signatures. LSTM and 1D-CNN achieved moderate performance (F1-scores of 0.870 and 0.854, respectively), while DGC and GCN lagged further behind (F1-scores of 0.797 and 0.806).

Overall, HSTGNN consistently achieves the highest detection accuracy and robustness across all fault types. Its superior AUC values across Tables 2-4 (ranging from 0.907 to 0.989) demonstrate stability across different classification thresholds. The results clearly indicate that HSTGNN's ability to jointly model complex spatial graph structures and long-term temporal dependencies allows it to detect both abrupt (bias) and gradual (drift) faults, as well as their combinations, while minimizing false positives.

In addition to Precision, Recall, and F1-score, we evaluated the models using the ROC curve and the corresponding AUC. The ROC curve plots the True Positive Rate against the False Positive Rate at varying classification thresholds, providing a visual assessment of a model's ability to discriminate between normal and faulty readings. A higher AUC indicates stronger overall discriminative performance. Figure 3 shows ROC curves for the Bias, Drift, and Mixed fault scenarios, illustrating that HSTGNN consistently achieves the highest True Positive Rate across nearly all False Positive Rate levels in every scenario. This demonstrates its superior capability in detecting both abrupt and gradual faults. In contrast, baseline models exhibit lower ROC curves, particularly under Drift and Mixed scenarios, highlighting their reduced effectiveness in capturing subtle or overlapping anomalies. Overall, the figure reinforces that HSTGNN's integrated spatial-temporal modeling effectively captures complex anomaly patterns across diverse fault types.

Table 5. Fault detection performance (AUC) disaggregated by sensor modality and fault scenario.

Method	**Group**	**Bias AUC**	**Drift AUC**	**Mixed AUC**
LSTM	Flow	0.986	0.909	0.945
	Temp	0.936	0.838	0.881
1D-CNN	Flow	0.937	0.880	0.907
	Temp	0.948	0.841	0.887
DGC	Flow	0.984	0.927	0.957
	Temp	0.808	0.758	0.768
GCN	Flow	0.986	**0.943**	**0.965**
	Temp	0.813	0.763	0.773
HSTGNN	Flow	**0.988**	0.940	**0.965**
	Temp	**0.989**	**0.888**	**0.933**

The performance breakdown by sensor group, shown in Table 5, provides important insight into how well faults can be detected within the district heating network. Across all sensors, bias faults remain the easiest to detect, while drift faults are consistently more challenging, reflecting the difficulty of identifying gradual signal changes. The mixed fault scenario follows a similar trend, with performance generally lying between bias and drift results.

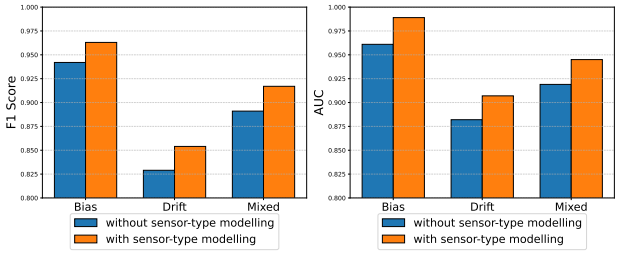

Figure 4. Comparison of AUC and F1 for HSTGNN (with sensor-type modelling) and the STGNN (without sensor-type modelling) across bias, drift, and mixed fault scenarios.

A clear difference also appears between sensor types. Standard graph-based methods such as DGC and GCN perform strongly on flow sensors but experience a marked decline when applied to temperature sensors, with AUC scores decreasing by up to approximately 18%. This indicates that homogeneous graph convolutions have difficulty capturing the slower dynamics and high thermal inertia characteristic of temperature signals. This performance gap is consistent across all fault types, including the mixed scenario.

In contrast, the proposed HSTGNN delivers strong and consistent performance across both sensor groups and all fault types. Notably, it eliminates the performance drop observed for temperature sensors, achieving an AUC of 0.989 for bias and 0.933 for mixed faults, compared to 0.813 and 0.773 for GCN, respectively.

5.4. Ablation Study

To evaluate the specific contribution of the proposed sensor-type modelling component, we conducted an ablation study. We compared our full HSTGNN architecture against a configuration where sensor-type modelling is omitted. In this simplified version, all sensor data are fed directly into a standard STGNN using a single, shared processing branch for all sensors, regardless of their modality.

As illustrated in Figure 4, the inclusion of sensor-type modelling leads to a significant increase in detection performance. Specifically, the AUC scores across all fault scenarios—bias, drift, and mixed—demonstrate that explicitly accounting for node-level heterogeneity allows the model to better capture the unique error characteristics of different sensor types compared to a homogeneous modelling approach.

6. CONCLUSION

This work introduced HSTGNN, a framework for virtual sensing–based sensor quality monitoring in district heating systems. The proposed model generates reference estimates for sensor measurements and enables the detection of sensor faults by identifying deviations between observed measurements and virtual sensor predictions. The architecture employs a multi-stream design that captures the heterogeneous temporal dynamics associated with different sensor types, learning spatial dependencies between sensors through a latent graph learning mechanism. A cross-type attention module further integrates information across sensor types, allowing the model to capture interactions between thermal and hydraulic measurements. Experimental results on simulated bias and drift scenarios demonstrate that the proposed approach can effectively detect sensor anomalies by leveraging both temporal dynamics and spatial relationships within the network. These findings highlight the potential of combining sensor-type–specific modeling with data-driven spatial learning to improve the reliability of sensor quality monitoring in district heating infrastructure. Future work will focus on evaluating the framework on larger real-world deployments and extending the approach toward more comprehensive sensor diagnostics, including fault localization and root-cause analysis.

ACKNOWLEDGMENT

This research was funded by the Swiss Federal Institute of Metrology (METAS). The authors would like to thank Kamstrup for providing the smart heat meters used in this study. We also gratefully acknowledge the Aalborg Smart Water Infrastructure Laboratory for their support and for providing the experimental platform used in the evaluation.

REFERENCES

Atwood, J., & Towsley, D. (2016). Diffusion-convolutional neural networks. *Advances in neural information processing systems*, *29*.

Bank, T., Madsen, F. W., Mortensen, L. K., Søndergaard, H. A. N., & Shaker, H. R. (2023). Virtual sensor-based fault detection and diagnosis framework for district heating systems: A top-down approach for quick fault localisation. In *Energy informatics academy conference* (pp. 292–307).

Belgacem, H., & Chihi, I. (2025). Toward reliable and intelligent sensor systems: A comprehensive study of fault diagnosis and mitigation. *IEEE Sensors Reviews*.

Bengio, Y., Léonard, N., & Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Cini, A., Zambon, D., & Alippi, C. (2023). Sparse graph learning from spatiotemporal time series. *Journal of Machine Learning Research*, *24*(242), 1–36.

Darvishi, H., Ciuonzo, D., Eide, E. R., & Rossi, P. S. (2020). Sensor-fault detection, isolation and accommodation for digital twins via modular data-driven architecture. *IEEE Sensors Journal*, *21*(4), 4827–4838.

Fink, O., Nejjar, I., Sharma, V., Faghih Niresi, K., Sun, H., Dong, H., ... Kesmen, Y. (2026). From physics to

machine learning and back: Part ii - learning and observational bias in prognostics and health management (phm). *Reliability Engineering & System Safety*, *274*, 112376.

Fink, O., Sharma, V., Nejjar, I., Von Krannichfeldt, L., Garmaev, S., Zhang, Z., ... Steiner, K. (2026). From physics to machine learning and back: Part i - learning with inductive biases in prognostics and health management (phm). *Reliability Engineering & System Safety*, *271*, 112213.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hsu, C.-C., Frusque, G., & Fink, O. (2023). A comparison of residual-based methods on fault detection. In *Annual conference of the phm society* (Vol. 15).

Huang, P., Copertaro, B., Zhang, X., Shen, J., Löfgren, I., Rönnelid, M., ... Svanfeldt, M. (2020). A review of data centers as prosumers in district energy systems: Renewable energy integration and waste heat reuse for district heating. *Applied energy*, *258*, 114109.

Jin, M., Koh, H. Y., Wen, Q., Zambon, D., Alippi, C., Webb, G. I., ... Pan, S. (2024). A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, *46*(12), 10466–10485.

Kamstrup. (2026). *MULTICAL® 303*. Retrieved from `https://www.kamstrup.com/en-en/product-centre/multical-303` (Accessed: 2026-03-23)

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kool, W., Van Hoof, H., & Welling, M. (2020). Ancestral gumbel-top-k sampling for sampling without replacement. *Journal of Machine Learning Research*, *21*(47), 1–36.

Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*.

Niresi, K. F., Bissig, H., Baumann, H., & Fink, O. (2024). Physics-enhanced graph neural networks for soft sensing in industrial internet of things. *IEEE Internet of Things Journal*, *11*(21), 34978–34990.

Niresi, K. F., Jensen, C. M., Kallesøe, C. S., Wisniewski, R., & Fink, O. (2026). Virtual smart metering in district heating networks via heterogeneous spatial-temporal graph neural networks. *arXiv preprint arXiv:2604.10166*.

Niresi, K. F., Kuhn, L., Frusque, G., & Fink, O. (2024). Informed graph learning by domain knowledge injection and smooth graph signal representation. In *2024 32nd european signal processing conference (eusipco)* (pp. 2467–2471).

Niresi, K. F., Nejjar, I., & Fink, O. (2025). Efficient unsupervised domain adaptation regression for spatial-temporal sensor fusion. *IEEE Internet of Things Journal*.

Sun, Q., & Ge, Z. (2021). A survey on deep learning for data-driven soft sensors. *IEEE Transactions on Industrial Informatics*, *17*(9), 5853–5866.

Theiler, R., & Fink, O. (2025). Heterogeneous graph neural networks for short-term state forecasting in power systems across domains and time scales: A hydroelectric power plant case study. *arXiv preprint arXiv:2507.06694*.

Val Ledesma, J., Wisniewski, R., & Kallesøe, C. S. (2021). Smart water infrastructures laboratory: Reconfigurable test-beds for research in water infrastructures management. *Water*, *13*(13), 1875.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, *32*(1), 4–24.

Yoon, S., Choi, Y., Koo, J., Hong, Y., Kim, R., & Kim, J. (2020). Virtual sensors for estimating district heating energy consumption under sensor absences in a residential building. *Energies*, *13*(22), 6013.

Zhao, M., Taal, C., Baggerohr, S., & Fink, O. (2025). Graph neural networks for virtual sensing in complex systems: Addressing heterogeneous temporal dynamics. *Mechanical Systems and Signal Processing*, *230*, 112544.