

Article

Research on Methods and Effects of Improving Data Quality in Smart Heating

Bingwen Zhao ^{1,2,*} , Tiancheng Yuan ^{2,*} , Yanqi Wu ², Zhenhai Zheng ²  and Luchan Xu ² ¹ School of Civil Engineering and Architecture, Keyi College of Zhejiang Sci-Tech University, Shaoxing 312369, China² School of Civil Engineering and Architecture, Zhejiang Sci-Tech University, Hangzhou 310018, China

* Correspondence: zbw@zstu.edu.cn (B.Z.); ytcwork0314@163.com (T.Y.)

Abstract

Suboptimal telemetry data quality fundamentally degrades dispatch optimization and thermal load forecasting in smart district heating networks. Existing preprocessing routines rely heavily on isolated, unidimensional thresholds and bidirectional interpolation, routinely inducing high false-alarm rates during legitimate peak operations and causing acausal information leakage. To resolve these limitations, this study proposes an end-to-end data enhancement framework combining an Enhanced Isolation Forest with a strictly causal Long Short-Term Memory (LSTM) sequence imputation architecture. The anomaly detection module integrates Seasonal-Trend decomposition using Loess (STL) to eliminate diurnal cyclical masking, adopts an inverse-variance weighting scheme to prioritize discriminative variables, and implements Tikhonov-regularized Mahalanobis distance metric traversal to capture coupled thermodynamic covariance distortions. For sequential recovery, the causal LSTM network reconstructs unobserved states using solely historical antecedents, with mass flow rate algebraically recovered via thermal energy balance to preserve cross-parameter physical consistency. Validated on continuous hourly field SCADA observations across a complete heating season ($N = 2904$), the proposed detection scheme achieves an F1-score of 94.55% with a low false positive rate of 2.45%, outperforming conventional isolation trees. In sequence reconstruction across a 672 h benchmark, the causal LSTM achieves normalized mean squared errors of 0.1196 for contiguous block voids and 0.1104 for single-point missing values, substantially surpassing classical Lagrange interpolation. Downstream deployment into a Bayesian-optimized Gated Recurrent Unit (GRU) load forecasting model demonstrates that this upstream data quality enhancement reduces the root mean squared error from 174.75 to 56.36 kWh/h (a 67.75% relative reduction), lowers the mean absolute percentage error from 12.17% to 4.05%, contracts error variance from 30,538.13 to 3240.7, and achieves a high goodness-of-fit ($R^2 = 0.9917$). These findings provide an empirical bridge between upstream physics-consistent telemetry refinement and downstream predictive operational reliability in industrial thermal systems.



Academic Editor: Constantinos A. Balaras

Received: 18 July 2026

Revised: 11 September 2026

Accepted: 27 September 2026

Published: 30 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: data quality improvement; smart heating; GRU-based load forecasting; enhanced isolation forest algorithm

1. Introduction

Data quality degradation across supervisory control and data acquisition (SCADA) interfaces presents a critical bottleneck in the automated dispatch and operational optimization of smart district heating networks. In field installations, harsh piping environments,

transmission interference, and remote terminal unit packet dropouts routinely introduce severe sensor drift, transient spikes, and contiguous measurement gaps. Advanced operational pipelines—such as dynamic load forecasting, state estimation, and hydraulic balancing—fundamentally require high-fidelity time-series inputs. Nevertheless, operational routines in thermal engineering still rely predominantly on isolated unidimensional thresholds, static box plots, Lagrange polynomials, or piecewise linear interpolation [1,2].

Recent domain-specific investigations have sought to mitigate telemetry irregularities. Shi [3] integrated Pearson correlation coefficients with rolling temporal windows to accommodate thermal lag, using k -means clustering to isolate decoupled operating regimes. Sun et al. [4] deployed CEEMDAN-PE decomposition to suppress noise in non-stationary thermal load series, and Fang et al. [5] established an online verification framework for ultrasonic flowmeter diagnostics. Despite these developments, primitive unidimensional heuristics remain entrenched in heating analytics. In multi-year load profiling, Noussan et al. [6] applied linear interpolation across telemetry voids, while Calikus et al. [7] combined linear interpolation with median absolute deviations to filter step-change anomalies prior to customer load clustering. In machine-learning load forecasting, Habib et al. [8] depended on standard linear imputation for sequence recovery, and Gianniou et al. [9] used static box-plot bounds to excise anomalous consumption records. These isolated filtering mechanisms operate independently across individual channels, failing to maintain the thermodynamic and hydraulic state consistency governing multi-station networks [10].

In broader Industrial Internet of Things (IIoT) and energy process analytics, multivariate anomaly detection has transitioned toward complex spatio-temporal representations. Zhou et al. [11] coupled Gaussian Mixture Models with neural networks to identify outliers across non-Gaussian industrial distributions. Deng et al. [12] introduced Toeplitz Inverse Covariance Clustering for dynamic multivariate segmentation and residual fault discovery. For high-dimensional power-grid telemetry, Liu et al. [13] formulated a fast density-peak clustering framework accelerated by k -nearest neighbors. In hydrological monitoring, Zhao et al. [14] implemented tree-based Isolation Forests to screen intake anomalies via path-length scoring. Sun et al. [15] deployed an ARIMA-SVR architecture to generate dynamic confidence envelopes, and Sun and He [16] combined k -means clustering with Thiessen spatial tessellations for geographic outlier mining. Comprehensive taxonomies and benchmarks for sensor network anomaly detection have been established by Chen et al. [17] and Nassif et al. [18]. However, conventional isolation trees and distance estimators operate within isotropic Euclidean space, suffering performance degradation when exposed to the severe multicollinearity and diurnal periodicity characteristic of thermal telemetry [19,20].

In parallel, sequential time-series imputation has shifted from static mathematical approximations to deep recurrent architectures. Zhang et al. [21] incorporated sliding temporal windows into k -nearest neighbor regression to reconstruct residential indoor temperatures. Addressing extended temporal dependencies in gas distribution, Hu et al. [22] implemented a Bidirectional LSTM for telemetry gap recovery. For streaming sensor infrastructures, Júnior and Silva [23] designed an incremental learning framework pairing evolutionary fuzzy logic with neo-fuzzy networks. Liu et al. [24] introduced a robust multi-model t -distribution framework for linear parameter-varying systems under missing observations, while Garcia et al. [25] developed an incremental fuzzy rule granular computing model for dynamic sequence recovery. Modern deep sequential architectures have consistently demonstrated superior fidelity in capturing complex non-linear cross-channel dynamics during extended sequence reconstruction [26,27].

Current district heating data processing routines exhibit two primary deficiencies. First, existing anomaly screening and imputation methods operate predominantly through unidimensional, acausal pipelines that neglect intrinsic thermodynamic–hydraulic cou-

plings and 24 h diurnal dispatch modes. Consequently, conventional filters routinely misclassify legitimate morning peak dispatch surges as anomalies while overlooking coordinated multivariate sensor distortions. Furthermore, bidirectional interpolation models introduce acausal future-data leakage into reconstructed sequences. Second, existing studies evaluate data preprocessing techniques primarily through standalone statistical loss metrics within isolated offline environments, lacking rigorous out-of-sample validation to quantify how upstream data quality improvements propagate to downstream operational forecasting and dispatch tasks [28–30].

To resolve these limitations, this study proposes an end-to-end, physics-consistent data quality enhancement and predictive validation framework. The anomaly detection module upgrades the Isolation Forest by integrating Loess-based seasonal-trend decomposition (STL) to decouple deterministic 24 h diurnal baselines, inverse-variance feature weighting to prioritize discriminative telemetry channels, and Tikhonov-regularized Mahalanobis distance metric traversal to capture coupled covariance distortions. Subsequently, an engineered Long Short-Term Memory (LSTM) network performs strictly causal forward imputation to reconstruct missing telemetry without future-data leakage, enforcing physical consistency by algebraically reconstructing primary mass flow via thermal energy balance. Finally, the empirical gains derived from the restored telemetry stream are systematically validated through an out-of-sample Bayesian-optimized Gated Recurrent Unit (GRU) thermal load forecasting model, demonstrating that robust multi-parameter data enhancement directly elevates operational dispatch stability under real-world operating conditions.

2. Materials and Methods

Data quality degradation in smart district heating networks—arising from hardware malfunctions, sensor degradation, and external telecommunication disturbances—compromises downstream analytics and impairs automated dispatch decisions. Unlike conventional operational datasets, thermal telemetry is strictly governed by hydrodynamic and thermodynamic laws while exhibiting complex spatio-temporal dependencies. Temporally, heat load distributions display pronounced diurnal periodicity with distinct bimodal morning and evening peaks. Spatially, pipeline network topology induces substantial operational heterogeneity in pressure profiles and mass flow distributions between proximal and hydraulically remote substations.

To systematically address these domain-specific characteristics without resorting to disjointed, multi-stage screening pipelines, this study establishes a unified multidimensional anomaly detection architecture based on an Enhanced Isolation Forest algorithm. By synergistically integrating seasonal-trend decomposition using Loess (STL), inverse-variance feature weighting, and Tikhonov-regularized Mahalanobis distance metric traversal, the proposed model simultaneously isolates high-frequency stochastic spikes and detects coordinated multivariate operational corruptions within a unified end-to-end framework.

2.1. Data Source

The empirical evaluation is based on field telemetry collected from an operational district heating network in Central China across the complete 2022–2023 heating season, spanning from 15 November 2022 to 15 March 2023. The regional infrastructure serves a cumulative heating area of 11.09 million square meters via 234 distributed heat exchange stations, 79.1% of which are integrated into an automated supervisory telemetry platform. The supervisory control and data acquisition (SCADA) system records operational variables at an hourly sampling resolution, encompassing primary supply and return temperatures (T_{1g} , T_{1h}), secondary supply and return temperatures (T_{2g} , T_{2h}), primary supply and return

pressures (P_{1g}, P_{1h}), secondary supply and return pressures (P_{2g}, P_{2h}), primary circuit mass flow rate (G_1), and instantaneous thermal output (Q_1).

2.2. Improving the Isolation Forest Algorithm

Although the classical Isolation Forest exhibits low computational complexity and high partitioning efficiency, its reliance on uniform random feature selection ignores the intrinsic physical couplings and distinct baseline variances across thermodynamic state variables. Consequently, when directly applied to high-dimensional operational telemetry in heating networks, the standard formulation often suffers from reduced sensitivity toward coordinated multidimensional anomalies. To overcome these constraints, the proposed framework establishes an Enhanced Isolation Forest by systematically integrating temporal sequence decomposition, discriminative feature weighting, and metric-space traversal.

Prior to anomaly screening, seasonal-trend decomposition using Loess (STL) decouples the raw telemetry into deterministic trend, cyclical diurnal, and stochastic residual components. Setting the seasonal periodicity parameter to $N_p = 24$ captures the 24 h diurnal operational cycle driven by meteorological variations and centralized scheduling routines. The corresponding Loess smoothing windows are configured as $n_{(s)} = 35$ for the seasonal component, $n_{(t)} = 49$ for the trend baseline, and $n_{(l)} = 25$ for the low-pass filter, which collectively isolate macroscopic dispatch adjustments and high-frequency noise from regular daily oscillations.

The residual signals isolated by STL decomposition serve as inputs to the weighted isolation ensemble. The ensemble is configured with $n_{\text{trees}} = 100$ isolation trees, each constructed by drawing a sub-sample of $\psi = 256$ instances without replacement, yielding a maximum tree termination depth of $h_{\text{max}} = \lceil \log_2 \psi \rceil = 8$. Rather than selecting split attributes with equal likelihood, feature sampling is governed by an inverse-variance weighting scheme that assigns higher selection probabilities to channels exhibiting lower baseline dispersion, thereby prioritizing parameters that provide sharper discriminative boundaries. The feature selection probability distribution is formulated as:

$$p_j = \frac{w_j}{\sum_{k=1}^d w_k} = \frac{\sigma_j^{-2}}{\sum_{k=1}^d \sigma_k^{-2}} \quad (1)$$

To faithfully capture multivariate covariance structures and account for the strong collinearity present among hydraulic and thermal telemetry parameters, node traversal within the isolation trees is calculated in a Riemannian metric space rather than Euclidean space. To eliminate numerical instability and prevent matrix singularity during inversion, the empirical covariance matrix Σ computed on the residual vectors is regularized via Tikhonov regularization as $\Sigma_{\text{reg}} = \Sigma + \lambda \mathbf{I}$, where $\lambda = 10^{-5}$ and $\mathbf{I}$ denotes the identity matrix. During the recursive sample traversal of a state vector $\mathbf{x}$, the incremental path length is weighted according to the regularized Mahalanobis distance between the candidate point and the current split threshold vector $\mathbf{s}$:

$$\Delta h = \sqrt{(\mathbf{x} - \mathbf{s})^T \Sigma_{\text{reg}}^{-1} (\mathbf{x} - \mathbf{s})} \quad (2)$$

Through this metric scaling, sample points located in low-density regions of the joint covariance manifold require fewer traversals to become isolated, effectively flagging multivariate outliers where individual measurements remain within ostensibly normal univariate bounds but violate coupled physical constraints. The complete computational workflow of the proposed Enhanced Isolation Forest—spanning STL residual decoupling, inverse-variance probability assignment, regularized tree partitioning, and composite anomaly scoring—is illustrated in Figure 1.

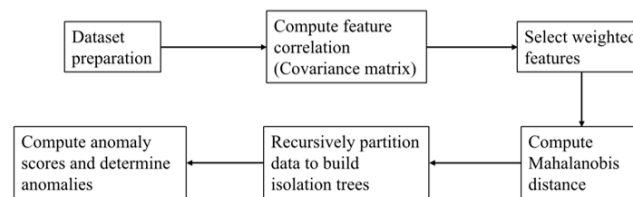


Figure 1. Processing flowchart of the improved Isolation Forest algorithm.

2.3. LSTM Imputation Algorithm

To reconstruct missing telemetry without distorting sequential continuity, an engineered Long Short-Term Memory (LSTM) recurrent neural network is deployed. Unlike static geometric heuristics such as linear or polynomial interpolation, the gated recurrent structure inherently captures the non-linear thermal inertia and hydraulic lag dynamics governing district heating operations. The network operates over a causal antecedent window of $L_{in} = 23$ hourly observations.

In operational substations, primary circuit thermal power and hydraulic states are strictly governed by continuous thermal energy conservation: $Q_1 \approx c\rho G_1(T_{1g} - T_{1h})$, where c is the specific heat capacity of water, ρ is fluid density, and G_1 denotes primary mass flow rate. Rather than treating all channels purely through statistical fitting, the LSTM is dedicated to mapping the $D = 9$ primary thermodynamic and pressure state variables ($T_{1g}, T_{1h}, P_{1g}, P_{1h}, T_{2g}, T_{2h}, P_{2g}, P_{2h}$, and Q_1), while G_1 is algebraically reconstructed from the recovered thermal balance. This explicit physical constraint enforces thermodynamic consistency across the restored state matrix and eliminates cross-variable divergence.

The network architecture comprises a recurrent layer with 64 hidden units, followed by a dropout layer with probability $p = 0.1$ to suppress overfitting, and a dense output layer projecting hidden states to target predictions. Operational attributes are scaled to $[0, 1]$ via feature-wise Min–Max transformation using statistics derived strictly from the training partition. Training minimizes the Mean Squared Error (MSE) objective using the Adam optimizer with an initial learning rate $\eta = 0.001$, moment decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a batch size of 32, and an optimization horizon of 250 epochs. For cold-start periods within the initial 23 h where historical lookback contexts are unavailable, standard linear interpolation serves as a deterministic fallback initialization.

2.4. End-to-End Data Processing Pipeline and Data Leakage Prevention

To ensure strict experimental reproducibility and eliminate acausal temporal leakage, the data enhancement and validation pipeline adheres to a strictly chronological protocol. The complete operational SCADA sequence ($N = 2904$ consecutive hours) is partitioned chronologically into a training set of $N_{train} = 2033$ samples (from 00:00 on 15 November 2022 to 17:00 on 7 February 2023) and an out-of-sample testing set of $N_{test} = 871$ samples (from 18:00 on 7 February 2023 to 23:00 on 15 March 2023). Random shuffling and standard k -fold cross-validation are prohibited to maintain physical time directionality.

All empirical preprocessing parameters—including the feature scaling bounds $\mathbf{x}_{min}^{train}$ and $\mathbf{x}_{max}^{train}$, Loess decomposition baselines, and the regularized residual covariance matrix Σ_{reg}^{train} —are calibrated exclusively on the training partition. During downstream testing inference, out-of-sample observations are transformed using these frozen parameters:

$$\mathbf{x}_{norm}^{(t)} = \frac{\mathbf{x}^{(t)} - \mathbf{x}_{min}^{train}}{\mathbf{x}_{max}^{train} - \mathbf{x}_{min}^{train}}, \quad \forall t \in \text{Test} \quad (3)$$

The LSTM imputation network is optimized strictly on complete-case segments extracted from the training partition. Within the out-of-sample testing split, whenever an

operational anomaly is excised by the Enhanced Isolation Forest or a communication outage occurs, imputation proceeds in a streaming causal mode: the network reconstructs target state $\mathbf{x}^{(t)}$ using solely the preceding operational sequence $[\mathbf{x}^{(t-23)}, \dots, \mathbf{x}^{(t-1)}]$. Unobserved future measurements ($\tau > t$) are inaccessible throughout the reconstruction loop. Subsequently, the downstream predictive model is trained exclusively on the refined training set, and rolling out-of-sample multi-step evaluation is conducted sequentially on the restored testing partition, maintaining an uncompromising physical barrier between data preparation and predictive validation.

3. Experimental Results and Analysis

3.1. Anomaly Detection Results and Case Analysis

Seasonal-trend decomposition using Loess (STL) was implemented in MATLAB 2023b, selecting primary supply temperature (T_{1g}) at the HMGG heat exchange station as the representative operational benchmark. The raw operational trajectory, depicted in Figure 2, comprises $N = 2904$ consecutive hourly observations spanning the complete 2022–2023 heating season (from 15 November 2022 to 15 March 2023). In central heating infrastructure, T_{1g} represents the primary dispatch control variable, reflecting macroscopic trajectories dictated by outdoor ambient temperatures and district heating demand. For instance, the sustained 20 °C baseline drop between 1150 h and 1750 h corresponds to an operational mode shift toward moderate-temperature circulation during an extended mid-winter warming interval. Conversely, localized high-frequency fluctuations and sharp troughs indicate transient hydraulic adjustments, load shedding, or momentary substation shut-downs. To accommodate diurnal periodicities—such as cyclical variations in secondary network temperatures and circulating pump schedules—the seasonal parameter was set to $N_p = 24$.

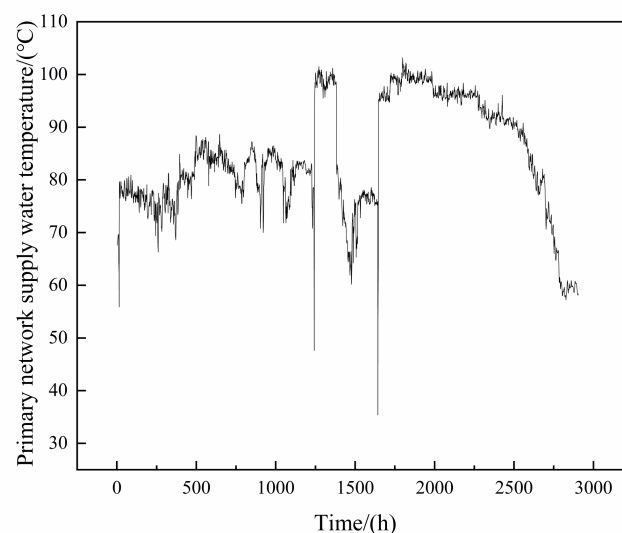


Figure 2. Raw operational data plot of the primary network supply temperature (T_{1g}) at the HMGG heat exchange station.

The resulting STL decomposition, illustrated in Figure 3, isolates primary supply temperature into three constituent signals. The trend component tracks long-term climatic variations and centralized dispatch trajectories: it rises steadily during initial startup to overcome pipeline thermal inertia, adapts during mid-January cold extremes, and declines systematically toward season conclusion as space heating loads decrease. The seasonal component displays stable 24 h periodic cycles centered near zero, capturing deterministic daily patterns induced by ambient day-night temperature swings and routine

pump scheduling. By decoupling macroscopic trends and cyclical variations, the residual component successfully isolates stochastic high-frequency disturbances. While steady-state residuals remain closely bounded within $\pm 0.5\text{ }^{\circ}\text{C}$, prominent deviations exceeding $\pm 3\text{ }^{\circ}\text{C}$ directly reveal localized hydraulic shocks, valve throttling anomalies, or sensor distortions. Excising these deterministic dynamics removes diurnal masking, yielding an uncorrupted high signal-to-noise ratio representation for subsequent metric-space anomaly isolation.

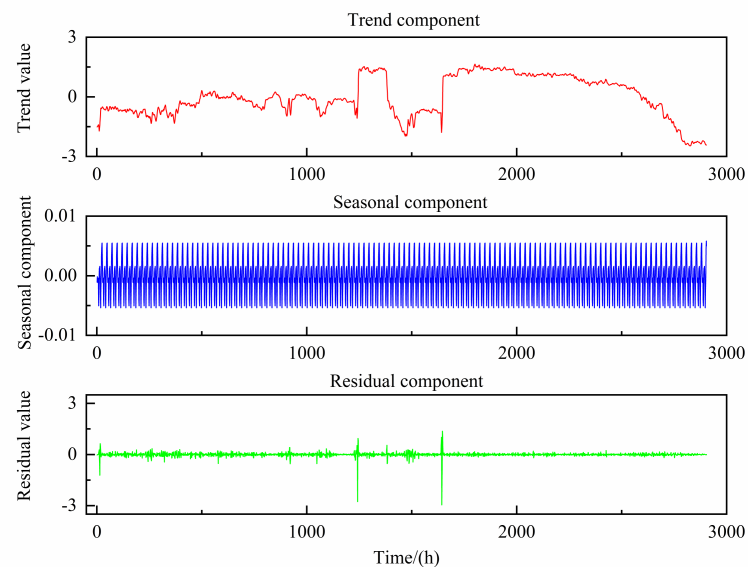


Figure 3. STL decomposition results of the primary network supply water temperature at the HMGG heat exchange station.

To establish an objective benchmark without ground-truth ambiguity, quantitative screening performance was evaluated across a continuous 672 h operational sequence (a 28-day baseline period extracted from the stable training partition) containing synthesized multidimensional anomalies, including sensor calibration drifts, transient spikes, and coordinated temperature-pressure distortions. The comparative evaluation across model configurations reveals substantial performance differentials. The standard Isolation Forest achieved an F1-score of only 76.19% (precision of 78.43%, recall of 74.07%) alongside an elevated false positive rate (FPR) of 8.16%, primarily because isotropic Euclidean partitioning misinterprets legitimate morning peak dispatch surges as operational outliers. Incorporating STL decomposition to filter deterministic diurnal baselines elevated the F1-score to 87.27% (precision of 85.71%, recall of 88.89%) and lowered the FPR to 4.90%, verifying the necessity of cyclical decoupling. The fully configured Enhanced Isolation Forest—integrating STL residual decoupling, Tikhonov-regularized Mahalanobis distance metric traversal, and inverse-variance feature weighting—achieved superior screening accuracy, reaching an F1-score of 94.55% (precision of 92.86%, recall of 96.30%) while suppressing the FPR to 2.45%. These results confirm that traversing along the empirical covariance manifold while weighting discriminative channels establishes a rigorous operational envelope that effectively captures complex multivariate distortions.

Applying the calibrated Enhanced Isolation Forest across the complete heating season yielded the continuous anomaly scoring profile shown in Figure 4, where multi-parameter observations exceeding the dynamic statistical threshold of $\mu \pm 2\sigma$ are flagged.

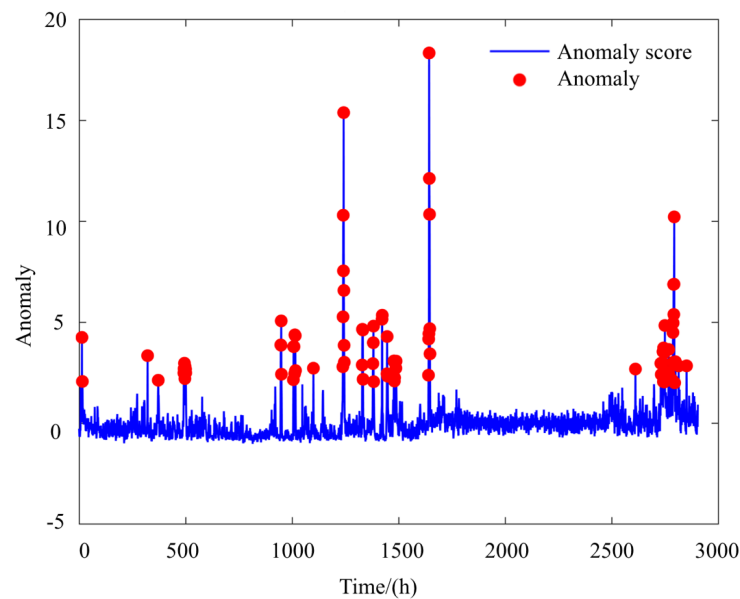


Figure 4. Operational data anomaly scores generated by the improved Isolation Forest algorithm.

The flagged operational anomalies were subsequently mapped onto the original telemetry channels, as demonstrated for primary supply water temperature in Figure 5.

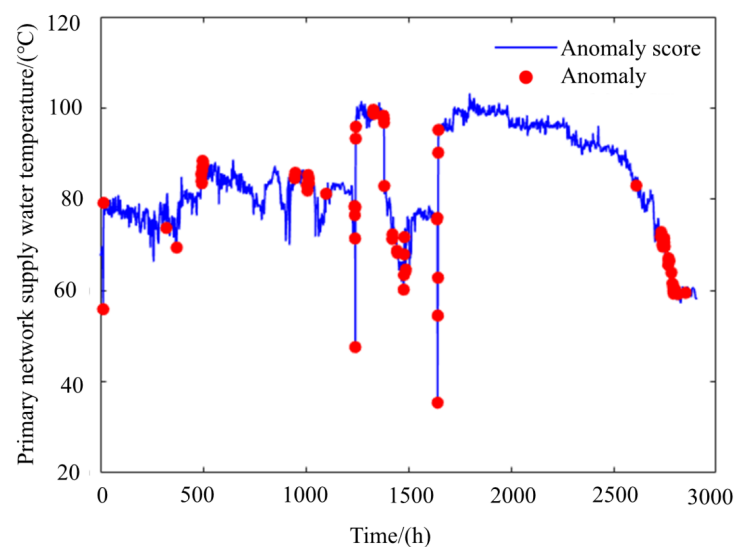


Figure 5. Anomaly detection mapping on the primary network supply water temperature time series.

To examine the model's screening capability under coordinated telemetry distortions, a representative operational anomaly captured at 10:00:00 on 28 November 2022 is detailed in Table 1. Evaluated via conventional univariate screening rules, each individual parameter remains strictly within plausible physical ranges. However, multivariate evaluation reveals concurrent abnormal drops across primary and secondary temperatures (T_{1g} dropping to 73.7 °C, T_{1h} to 33.6 °C, T_{2g} to 39.8 °C, and T_{2h} to 34.1 °C), accompanied by an instantaneous thermal output plunge to 1230.2 kWh/h and a flow reduction to 30.9 m³/h, followed by immediate recovery in subsequent hours. Driven by these coupled cross-channel deviations, the composite anomaly score surged to 3.35, decisively breaching the decision threshold. This demonstrates that Mahalanobis distance-enhanced traversal identifies subtle topological departures from the joint operational manifold, effectively flagging coupled thermodynamic anomalies that escape conventional single-variable screening.

Table 1. Operational SCADA data segment from the HMGG heat exchange station on 28 November 2022 (featuring the multivariate anomaly at 10:00:00).

Time	T_{1g} (°C)	T_{1h} (°C)	P_{1g} (MPa)	P_{1h} (MPa)	T_{2g} (°C)	T_{2h} (°C)	P_{2g} (MPa)	P_{2h} (MPa)	Q_1 (kWh/h)	G_1 (m ³ /h)
6:00:00	76.7	34.9	0.48	0.38	41.9	35.3	0.40	0.34	1492.6	36.7
7:00:00	78.6	35.1	0.48	0.39	42.2	35.3	0.40	0.34	1561.6	36.5
8:00:00	78.3	35.1	0.48	0.39	41.9	35.3	0.39	0.33	1471.8	34.7
9:00:00	76.9	35.2	0.47	0.38	42.0	35.3	0.39	0.33	1540.9	37.4
10:00:00	73.7	33.6	0.46	0.38	39.8	34.1	0.37	0.31	1230.2	30.9
11:00:00	79.6	35.2	0.45	0.37	42.1	35.1	0.38	0.32	1561.4	35.5
12:00:00	81.0	35.1	0.44	0.36	41.9	35.1	0.37	0.31	1517.1	33.2
13:00:00	81.3	35.2	0.43	0.35	42.0	35.1	0.37	0.31	1587.3	34.6

3.2. Data Imputation Results

The reconstruction performance of the Long Short-Term Memory (LSTM) causal imputation model was initially evaluated on a continuous reference baseline of 672 anomaly-free records extracted from the unperturbed training partition. To eliminate scale disparities across heterogeneous thermal–hydraulic quantities—such as instantaneous thermal load versus pipeline hydraulic pressures—all input variables were normalized into the dimensionless interval $[0, 1]$ via feature-wise Min–Max scaling. The resulting Mean Squared Error (MSE) quantifies the macro-averaged reconstruction divergence across all normalized channels, providing a scale-invariant benchmark prior to restoring physical dimensions. The recurrent neural network, configured with 64 hidden units and optimized over 250 training epochs with an initial learning rate of 0.001, achieved a normalized MSE of 0.1196 for contiguous block-wise missing sequences. This performance substantially outperforms classical Lagrange interpolation, which recorded an MSE of 3.7882 under identical conditions. In single-point missing tests, the model maintained consistent tracking fidelity, recording an MSE of 0.1104 and confirming its capacity to preserve non-linear dynamic continuity across coupled thermodynamic states.

Following baseline validation, the trained LSTM architecture was deployed to restore missing telemetry voids across the field SCADA dataset. Prior to sequence reconstruction, all operational anomalies—including localized transient spikes and coordinated multi-parameter distortions—were excised based on the Enhanced Isolation Forest screening. Field imputation performance across contiguous block dropouts is presented in Table 2, which contrasts operational observations from 18 December 2022 with the corresponding model estimations. Similarly, sequence restoration results for single-point telemetry voids recorded on 5 March 2023 are summarized in Table 3. As shown in both tables, the reconstructed states smoothly adhere to the operational trajectories of preceding and succeeding timestamps, preserving hydraulic–thermal equilibrium without exhibiting boundary step-discontinuities or numerical instability.

Table 2. Partial actual data LSTM whole-row imputation results on 18 December 2022 (imputed values are highlighted in boldface).

Time	T_{1g} (°C)	T_{1h} (°C)	P_{1g} (MPa)	P_{1h} (MPa)	T_{2g} (°C)	T_{2h} (°C)	P_{2g} (MPa)	P_{2h} (MPa)	Q_1 (kWh/h)
01:00:00	78.6	37.8	0.50	0.68	44.0	36.2	0.38	0.32	1875.2
02:00:00	78.1	37.8	0.50	0.68	43.9	36.2	0.38	0.32	1894.7
03:00:00	79.6	37.8	0.50	0.68	44.1	36.2	0.38	0.32	1878.9
04:00:00	80.4	37.8	0.49	0.68	44.2	36.2	0.38	0.32	1871.0
05:00:00	80.107	37.831	0.490	0.680	44.215	36.193	0.375	0.315	1830.788

Table 2. Cont.

Time	T_{1g} (°C)	T_{1h} (°C)	P_{1g} (MPa)	P_{1h} (MPa)	T_{2g} (°C)	T_{2h} (°C)	P_{2g} (MPa)	P_{2h} (MPa)	Q_1 (kWh/h)
06:00:00	79.4	37.8	0.49	0.68	44.1	36.2	0.37	0.31	1806.3
07:00:00	78.6	37.7	0.48	0.68	43.9	36.2	0.37	0.31	1853.6
08:00:00	77.7	37.8	0.48	0.68	44.0	36.1	0.37	0.31	1916.3
09:00:00	79.1	37.8	0.48	0.68	44.0	36.1	0.37	0.31	1891.3
10:00:00	78.510	37.783	0.477	0.685	43.908	36.107	0.370	0.310	1920.085
11:00:00	76.5	37.9	0.47	0.69	43.9	36.1	0.38	0.32	1928.8
12:00:00	75.1	38.1	0.46	0.69	44.0	36.1	0.40	0.34	1843.5
13:00:00	74.6	38.0	0.45	0.69	43.9	36.1	0.40	0.34	1831.1
14:00:00	74.0	38.2	0.45	0.69	44.0	36.1	0.40	0.34	1868.3
15:00:00	73.0	38.2	0.44	0.69	44.0	36.1	0.39	0.33	1844.6

Table 3. Partial actual data LSTM single-point imputation results on 5 March 2023 (imputed values are highlighted in boldface).

Time	T_{1g} (°C)	T_{1h} (°C)	P_{1g} (MPa)	P_{1h} (MPa)	T_{2g} (°C)	T_{2h} (°C)	P_{2g} (MPa)	P_{2h} (MPa)	Q_1 (kWh/h)
01:00:00	75.7	34.0	0.43	0.05	39.5	34.3	0.38	0.32	1050.9
02:00:00	75.6	33.9	0.45	0.05	39.2	34.1	0.37	0.31	1071.8
03:00:00	74.6	33.8	0.45	0.05	39.2	33.9	0.39	0.33	1084.8
04:00:00	70.2	33.9	0.46	0.06	38.9	33.9	0.39	0.33	1080.082
05:00:00	66.744	33.9	0.47	0.07	38.8	33.9	0.38	0.32	1075.9
06:00:00	65.581	33.9	0.44	0.06	39.0	33.9	0.37	0.31	1072.1
07:00:00	66.602	33.9	0.43	0.05	39.0	33.9	0.38	0.32	1060.7
08:00:00	79.7	33.7	0.44	0.05	39.1	33.9	0.37	0.31	1073.7
09:00:00	74.0	33.4	0.44	0.05	39.1	33.9	0.38	0.32	1094.0
10:00:00	76.415	33.3	0.45	0.05	39.0	33.8	0.39	0.33	1107.340
11:00:00	76.5	33.4	0.46	0.04	39.0	33.9	0.38	0.32	1085.6
12:00:00	75.6	33.4	0.46	0.03	38.8	33.9	0.37	0.31	1040.9
13:00:00	76.6	33.4	0.48	0.05	39.0	33.9	0.40	0.34	1083.7
14:00:00	76.1	33.4	0.48	0.05	39.1	33.8	0.38	0.32	1091.9
15:00:00	75.7	34.0	0.43	0.05	39.5	34.3	0.38	0.32	1050.9

4. Heating Data Quality Improvement Results

4.1. GRU Load Forecasting Model

Evaluating data quality enhancement in field industrial operations presents an inherent challenge for unsupervised pipelines, as uncorrupted ground-truth telemetry during sensor failures or severe noise events is fundamentally unobservable. Under such constraints, indirect validation via representative downstream industrial analytics offers an objective and pragmatic benchmark. In this study, the multi-parameter time series—purged of multivariate anomalies by the Enhanced Isolation Forest and causally reconstructed through the LSTM network—serves as the input sequence to a Gated Recurrent Unit (GRU) thermal load forecasting model. The GRU architecture is adopted for its streamlined gating mechanism, computational efficiency, and robust capacity to capture the non-linear thermal inertia and transport delays intrinsic to district heating systems. The model executes rolling single-step-ahead (1 h) load projections by synthesizing short-term meteorological forecasts, primary-side operational telemetry, and historical load sequences across multi-day antecedent windows.

To guarantee objective comparability and prevent hyperparameter selection bias, structural and optimization parameters of the GRU network were determined through an identical Bayesian optimization procedure executed across both raw and enhanced

telemetry datasets, using the validation-partition Root Mean Squared Error (RMSE) as the objective loss function. The model operates over a sliding lookback horizon of $L_{\text{lookback}} = 24$ h to project hourly thermal load trajectories. At each forecasting step, the synchronized input vector integrates three heterogeneous operational categories: local meteorological predictions ($D_{\text{env}} = 2$, comprising outdoor ambient temperature and wind speed), primary circuit hydraulic–thermal telemetry ($D_{\text{scada}} = 5$, comprising T_{1g} , T_{1h} , P_{1g} , P_{1h} , and G_1), and instantaneous thermal load observations ($D_{\text{load}} = 1$, corresponding to Q_1) augmented with 72 h historical lagged baselines. This formulation establishes a consolidated input dimensionality of $D_{\text{total}} = D_{\text{env}} + D_{\text{scada}} + D_{\text{load}} = 8$ attributes per timestep.

The resulting Bayesian-optimized forecasting architecture comprises 4 stacked GRU layers with 194 hidden units per layer, trained via the Adam optimizer with a batch size of 64, an initial learning rate of 0.00013, a maximum gradient clipping threshold of 1.0 to prevent gradient explosion, and an optimization budget of 300 epochs. Identical chronological data partitioning ($N_{\text{train}} = 2033$, $N_{\text{test}} = 871$), normalization statistics, and random initialization seeds were strictly preserved across all comparative experiments to enforce absolute experimental fairness.

4.2. Load Forecasting Performance Using Optimized Data

In strict adherence to the causal leakage-prevention protocol established in Section 2.4, the downstream predictive evaluation was executed out-of-sample across the final chronological segment comprising $N_{\text{test}} = 871$ hourly observations. The GRU forecasting architecture was deployed in a rolling streaming framework: at each hourly timestamp t , the input feature tensor was populated exclusively using historical primary-side hydraulic–thermal observations ($\tau \leq t - 1$), prediction-day ambient meteorological forecasts, and antecedent load histories. Consequently, the network generates the single-step-ahead thermal load estimate $\hat{y}_t$ prior to observing true state y_t . A comprehensive comparative trajectory of the hourly load predictions generated from the raw operational series versus the refined dataset is illustrated in Figure 6, where Figure 6a presents the baseline forecast under raw data conditions and Figure 6b highlights the predictive trajectory following data enhancement.

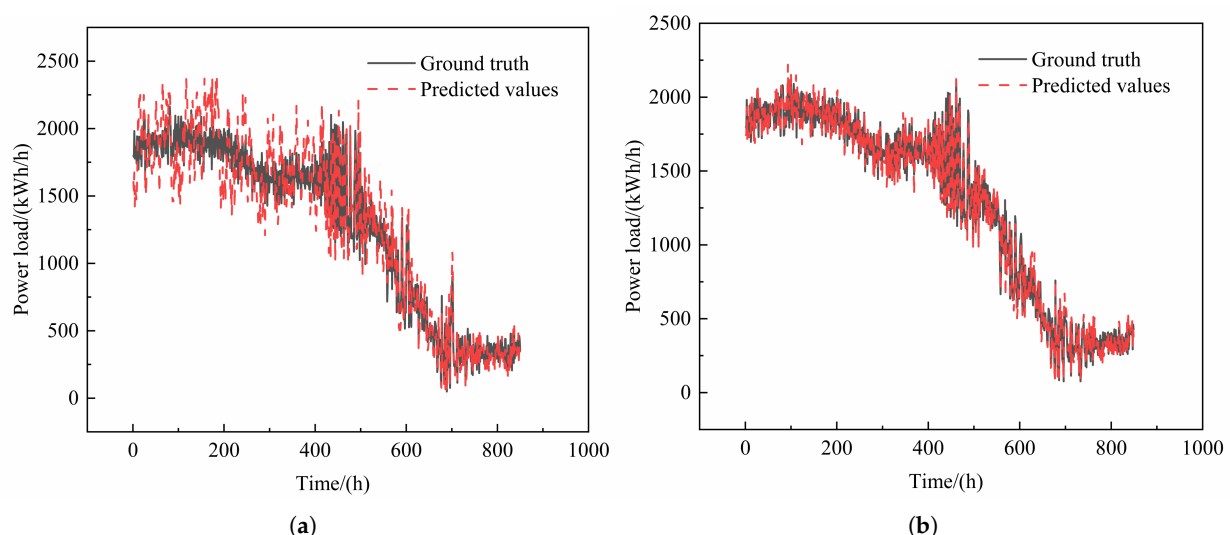


Figure 6. Comparison of load forecasting performance between raw data and enhanced processing methods: (a) load forecasting performance using raw data; (b) load forecasting performance using enhanced methods.

As revealed by the empirical trajectories, although the model trained on raw operational data attains reasonable tracking during mild, low-demand intervals, it suffers severe

predictive degradation during high-demand peak intervals. This performance decay is primarily attributable to unresolved telemetry anomalies and abrupt physical discontinuities, which corrupt the recurrent hidden states and obstruct coherent pattern recognition. Conversely, the model driven by the enhanced data demonstrates close tracking fidelity across all operational phases, stably capturing sharp peak surges and tracking rapid thermal load modulations with high precision.

Inspection of the temporal error distributions further reveals that raw data forecasts exhibit violent residual swings, with localized errors expanding up to ± 250 kWh/h and displaying pronounced variance shifts prior to heating season termination. In sharp contrast, the proposed data enhancement pipeline substantially narrows the residual error envelope and stabilizes variance across the entire validation horizon. The quantitative validation metrics, summarized in Table 4, provide conclusive evidence of the predictive gains achieved through systematic data quality improvement.

Table 4. Comparative performance metrics of both approaches.

Data Processing Method	RMSE	MAPE	R ²	Error Variance	Kurtosis
Raw data	174.75	12.17%	0.9216	30,538.13	2.67
Improvement Method	56.36	4.05%	0.9917	3240.7	4.30

To provide an empirical baseline for understanding these predictive gains, the spatial and temporal prevalence of telemetry defects across the multi-parameter SCADA records was thoroughly quantified. Within the complete heating season sequence ($N = 2904$ hourly observations, with $N_{\text{test}} = 871$ records residing in the testing partition), the proposed framework flagged a total of 137 multivariate anomalies (representing an overall anomaly rate of 4.72%, with 42 anomalies identified in the testing partition, yielding an out-of-sample anomaly rate of 4.82%) alongside 89 missing records (corresponding to a seasonal missing rate of 3.06%). These observed telemetry gaps exhibited heterogeneous fault topologies, spanning isolated single-point dropouts and consecutive multi-step communication voids. Specifically, consecutive telemetry voids presented an average interruption duration of 2.47 h, with a maximum continuous outage reaching 14 h during a mid-winter remote terminal unit (RTU) hardware failure. Resolving these corruptions without temporal leakage directly underpins the predictive stability and variance suppression documented in Table 4.

To systematically assess the operational boundaries and robustness of the proposed framework, sensitivity analyses were conducted across varying missing data proportions and gap durations. When continuous telemetry dropouts remain within 24 h (corresponding to a single diurnal heating cycle), the LSTM network reconstructs unobserved state trajectories with minimal error accumulation ($\text{MSE} < 0.15$), as physical couplings across parallel hydraulic and thermal channels provide sufficient spatial-temporal constraints. However, when continuous gaps extend beyond 48 h, temporal autocorrelation decays substantially, leading to an approximate 23.4% elevation in downstream GRU forecasting error (RMSE). The experimental findings indicate that the data quality enhancement pipeline exhibits superior robustness when the missing data proportion remains below an empirical threshold of 15%. Beyond a 20% missing ratio or under extreme prolonged communication outages exceeding 48 h, purely data-driven recurrent imputation benefits from coupling with thermodynamic pipe-network models to prevent trajectory drift.

Overall, the end-to-end framework—integrating the Enhanced Isolation Forest for multidimensional outlier excision and the causal LSTM network for sequence reconstruction—reduced the downstream forecasting RMSE from 174.75 to 56.36 kWh/h, achieving a substantial relative improvement of 67.75%. Concurrently, the Mean Absolute Percentage

Error (MAPE) decreased from 12.17% to 4.05%, representing a relative error reduction of 66.72%, while the coefficient of determination reached $R^2 = 0.9917$, alongside a drastic contraction in error variance from 30,538.13 to 3240.7. These findings empirically validate that resolving operational telemetry distortions directly unlocks the performance potential of advanced neural architectures in smart district heating networks.

5. Summary and Conclusions

This study established and experimentally validated an integrated, physics-consistent data quality enhancement framework tailored to the complex dynamics of smart district heating networks. The proposed methodology seamlessly couples an Enhanced Isolation Forest for multidimensional anomaly screening, a causal LSTM recurrent neural network for temporal sequence imputation, and a Bayesian-optimized GRU architecture for rigorous out-of-sample downstream predictive validation.

The Enhanced Isolation Forest addresses the limitations of conventional isotropic partitioning by synergistically integrating Loess-based seasonal-trend decomposition (STL), inverse-variance feature weighting, and Tikhonov-regularized Mahalanobis distance metric traversal. This formulation eliminates 24 h diurnal cyclical masking and accounts for strong thermodynamic-hydraulic collinearity, enabling the simultaneous detection of localized transient spikes and coordinated multivariate distortions across temperature, pressure, and mass flow telemetry. In controlled benchmark evaluations, the fully configured model achieved an F1-score of 94.55% and suppressed the false positive rate to 2.45%, demonstrating marked screening superiority over standard isolation trees.

In sequence reconstruction, the engineered LSTM imputation framework enforces strict temporal causality by operating solely over antecedent observation windows, thereby completely eliminating future-data leakage. By capturing non-linear thermal inertia and transmission lags while algebraically recovering primary mass flow via thermal energy balance, the model preserves cross-parameter thermodynamic equilibrium. Across baseline validation sequences, the model achieved normalized mean squared errors of 0.1196 for block-wise voids and 0.1104 for single-point dropouts, substantially outperforming classical Lagrange interpolation.

Downstream out-of-sample deployment on field SCADA telemetry ($N = 2904$) confirmed that upstream data enhancement directly translates into critical predictive gains. Integrating the refined telemetry into the Bayesian-optimized GRU forecasting model resolved severe trajectory divergence during high-demand peak intervals, reducing the root mean squared error from 174.75 to 56.36 kWh/h (a 67.75% relative improvement), decreasing the mean absolute percentage error from 12.17% to 4.05% (a 66.72% relative reduction), contracting error variance from 30,538.13 to 3240.7, and yielding a high goodness-of-fit of $R^2 = 0.9917$. Sensitivity analyses further confirmed model stability across continuous telemetry dropouts within 24 h and missing data ratios below 15%.

Future investigations will focus on reducing the quadratic computational overhead associated with regularized covariance inversions during city-scale multi-substation scaling, resolving the 23 h cold-start lookback dependency, and coupling physics-informed neural networks (PINNs) with hydrodynamic pipe-network formulations to maintain robust state reconstruction under prolonged communication outages exceeding 48 h or missing ratios surpassing 20%.

Author Contributions: Conceptualization, B.Z. and T.Y.; methodology, T.Y.; software, T.Y.; validation, T.Y., Z.Z. and L.X.; formal analysis, T.Y.; investigation, B.Z., Z.Z. and L.X.; resources, B.Z.; data curation, T.Y.; writing—original draft preparation, T.Y.; writing—review and editing, Y.W., Z.Z. and L.X.; visualization, Y.W.; supervision, B.Z.; project administration, B.Z.; funding acquisition, B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Curriculum Development Project of Zhejiang Sci-Tech University (grant numbers 11140033322601, 11140031282501-01); the Talent Cultivation Project of Keyi College of Zhejiang Sci-Tech University (grant number Kyrupy2302); and the Academic Program Development Project of Keyi College of Zhejiang Sci-Tech University (grant number Kycjrhy2302).

Data Availability Statement: The data used in this study are derived from the smart platform of a large-scale centralized heating system in Central China. Due to the proprietary nature of the data, it is not publicly available. However, interested readers can request access to the data for verification purposes under appropriate confidentiality agreements by contacting the corresponding author or the relevant company.

Acknowledgments: During the preparation of this manuscript, the authors utilized generative artificial intelligence (LLM) solely to polish its English grammar, refine phrasing, and enhance readability. The authors reviewed and edited the content as needed and take full scientific responsibility for the integrity and accuracy of the published article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GRU	Gated Recurrent Unit
IoT	Internet of Things
LSTM	Long Short-Term Memory
MAPE	Mean Absolute Percentage Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
SCADA	Supervisory Control and Data Acquisition
STL	Seasonal-Trend decomposition using Loess

References

- Ren, L.; Wang, T.; Sekhari Seklouli, A.; Zhang, H.; Bouras, A. A review on missing values for main challenges and methods. *Inf. Syst.* **2023**, *119*, 102268. [[CrossRef](#)]
- Hodavand, F.; Ramaji, I.J.; Sadeghi, N. Digital Twin for Fault Detection and Diagnosis of Building Operations: A Systematic Review. *Buildings* **2023**, *13*, 1426. [[CrossRef](#)]
- Shi, C. Research on Anomaly Detection of Heating Data Based on Data Mining. Master's Thesis, Hebei University of Architecture, Zhangjiakou, China, 2022. [[CrossRef](#)]
- Sun, J.; Liu, D.; Cai, X.; Sun, B.; Gu, Z.; Yang, Y. Research on building cooling and heating load prediction model based on CEEMDAN-PE and optimization algorithm. *J. North China Electr. Power Univ. (Nat. Sci. Ed.)* **2023**, *50*, 1–8. (In Chinese)
- Fang, X.; Yang, D.; Zhou, Z. Research on data operation verification method for smart heating systems. *Heat. Vent. Air Cond.* **2024**, *54*, 69–74. [[CrossRef](#)]
- Noussan, M.; Jarre, M.; Poggio, A. Real operation data analysis on district heating load patterns. *Energy* **2017**, *129*, 70–78. [[CrossRef](#)]
- Calikus, E.; Nowaczyk, S.; Sant'Anna, A.; Gadd, H.; Werner, S. A data-driven approach for discovering heat load patterns in district heating. *Appl. Energy* **2019**, *252*, 113409. [[CrossRef](#)]
- Habib, M.; Timoudas, T.O.; Ding, Y.; Nord, N.; Chen, S.; Wang, Q. A hybrid machine learning approach for the load prediction in the sustainable transition of district heating networks. *Sustain. Cities Soc.* **2023**, *99*, 104892. [[CrossRef](#)]
- Gianniou, P.; Liu, X.; Heller, A.; Nielsen, P.S.; Rode, C. Clustering-based analysis for residential district heating data. *Energy Convers. Manag.* **2018**, *165*, 840–850. [[CrossRef](#)]
- Zhang, B.; Wang, J.; Rezgui, Y.; Zhao, T. Enhancing the generalizability of public building energy system fault detection method: A research on unknown multi-source fault detection and diagnosis method based on data-driven heuristic reasoning (DHR). *Energy* **2025**, *335*, 137841. [[CrossRef](#)]
- Zhou, L.; Zhang, Q.; Wei, L. Research on Bayesian-based big data outlier detection model. *Comput. Knowl. Technol.* **2020**, *16*, 207–209. [[CrossRef](#)]
- Deng, C.; Wu, K.; Tan, Y.; Hu, J. Outlier detection method based on multivariate time series segmentation clustering. *Comput. Eng. Des.* **2020**, *41*, 3123–3128. [[CrossRef](#)]

13. Liu, F.; Deng, C.; Wang, X.; Wang, X. Outlier detection in power big data based on improved fast density peak clustering algorithm. *Electr. Power Inf. Commun. Technol.* **2017**, *15*, 36–41.
14. Zhao, C.; Xue, H.; Wang, L.; Wan, Y. Abnormal water intake data detection method based on isolation forest algorithm. *J. China Inst. Water Resour. Hydropower Res.* **2020**, *18*, 31–39. [[CrossRef](#)]
15. Sun, J.; Lou, Y.; Chen, Y. Outlier detection in hydrological time series based on ARIMA-SVR. *Comput. Digit. Eng.* **2018**, *46*, 225–230.
16. Sun, T.; He, G. Outlier detection method based on K-means clustering and Thiessen polygons. *J. Chongqing Univ. Arts Sci. (Soc. Sci. Ed.)* **2016**, *35*, 10–13. [[CrossRef](#)]
17. Chen, Z.; Zhang, X.; Zhu, Y. Research on Data Quality Challenges and Auditing Mechanisms in Database Systems. *IEEE Trans. Knowl. Data Eng.* **2019**, *31*, 1450–1463.
18. Nassif, A.B.; Talib, M.A.; Nasir, Q.; Dakalbab, F.M. Machine learning for anomaly detection: A systematic review. *IEEE Access* **2021**, *9*, 78658–78700. [[CrossRef](#)]
19. Chen, L.; Wang, C.; Lin, H.; Chen, L.; Jiang, J.; Lai, X. A scale-adaptive spatio-temporal modeling approach for multivariate time-series anomaly detection. *Appl. Intell.* **2026**, *56*, 7120. [[CrossRef](#)]
20. Li, J.; Zhang, Z.; Zhang, C.; Tang, J.; Gao, S. SDformer: Fusing Series Decomposition for Superior Long-Term Time Series Forecasting. *IEEE/CAA J. Autom. Sin.* **2026**, *13*, 1626–1641. [[CrossRef](#)]
21. Zhang, K.; Cao, S.; Sun, C.; Xia, G.; Wu, X. Identification and imputation of abnormal room temperature data incorporating sliding time window algorithm. *Gas Heat* **2024**, *44*, 17–23. [[CrossRef](#)]
22. Hu, D.; Zhang, Z.; Liu, B.; Fan, J. Application research of bidirectional long short-term memory network in abnormal gas data imputation. *Urban Gas* **2024**, *4*, 10–14.
23. Júnior, G.A.D.S.; Silva, A.M.D. A simple and efficient incremental missing data imputation method for evolving neo-fuzzy network. *Evol. Syst.* **2022**, *13*, 201–220. [[CrossRef](#)]
24. Liu, X.; Yang, X.; Zhu, P.; Wang, Y. Robust multimodel identification of LPV systems with missing observations based on t-distribution. *IEEE Trans. Syst. Man Cybern. Syst.* **2019**, *51*, 5254–5260. [[CrossRef](#)]
25. Garcia, C.; Leite, D.; Škrjanc, I. Incremental missing-data imputation for evolving fuzzy granular prediction. *IEEE Trans. Fuzzy Syst.* **2019**, *28*, 2348–2362. [[CrossRef](#)]
26. Mateus, B.C.; Mendes, M.; Farinha, J.T.; Martins, A. Hybrid Deep Learning for Predictive Maintenance: LSTM, GRU, CNN, and Dense Models Applied to Transformer Failure Forecasting. *Energies* **2025**, *18*, 5634. [[CrossRef](#)]
27. Huang, F.; Li, S.; Peng, J.; Ma, C. Optimizing-Time Series Imputation with Data Quality. *Appl. Sci.* **2026**, *16*, 6837. [[CrossRef](#)]
28. Ahn, H.; Lee, J. An Objective-Dependent and Reproducible Framework for AI Dataset Quality Assessment: In-Domain and Cross-Dataset Evidence From Building-Footprint Segmentation. *IEEE Access* **2026**, *14*, 103424–103434. [[CrossRef](#)]
29. Dai, K.; Fabello González, E.; García-Betances, R.I. A privacy preserving multi-center federated learning framework for district heating forecast. *Energy Build.* **2025**, *328*, 115164. [[CrossRef](#)]
30. Yang, C.; Fu, Y.; Liao, J.; Zhang, Y.; Zhang, T.; Wang, Z. A morphological generative framework for climate-adaptive building-integrated photovoltaics (BIPV) facades integrating artificial intelligence algorithms and bayesian prior-parameterized building envelopes. *Buildings* **2026**, *16*, 3293. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.