DOI: https://doi.org/10.15276/aait.09.2026.26
UDC 004.94:664.723

Sensor data preprocessing pipeline for grain temperature monitoring and forecasting datasets in drying systems

Ihor I. Moroz[1)]
ORCID: https://orcid.org/0009-0004-9241-5859; moroz.i.i@vntu.edu.ua
Mariia S. Yukhymchuk[1)]
ORCID: https://orcid.org/0000-0002-8131-9739; umc1987@vntu.edu.ua. Scopus Author ID: 59159174300
[1)] Vinnytsia National Technical University, 95, Khmelnytskyi Hwy, Vinnytsia, 21021, Ukraine

ABSTRACT

Relevance. The quality of training data significantly affects the reliability of data-driven grain temperature forecasting systems. Sensor networks used in grain drying environments generate missing observations, measurement drift, and anomalous values caused by communication failures and sensor degradation. The absence of standardized preprocessing procedures reduces reproducibility and limits the practical applicability of predictive approaches. **Purpose.** The purpose of the study is to develop a methodology for sensor data collection and preprocessing intended for formation of datasets for grain temperature forecasting under continuous drying conditions. **Objectives.** The objectives include development of a multi-source monitoring architecture, justification of spatial sensor deployment, construction of a preprocessing pipeline for anomaly handling and missing value restoration, formation of feature vectors, and definition of a dataset partitioning strategy preserving temporal dependencies. **Methods.** The proposed methodology integrates digital temperature sensors, humidity sensors, carbon dioxide sensors based on non-dispersive infrared measurement, and meteorological data sources. Preprocessing consists of physical range validation, anomaly detection using the Isolation Forest algorithm, hierarchical missing value imputation, min-max normalization, cyclic temporal encoding, and chronological partitioning of datasets. **Scientific novelty** lies in the formalization of a topologic-spatiotemporal data validation framework that distinguishes local transducer degradation from collective thermal process transitions in non-stationary agricultural environments. This is achieved via a coupled multivariate Isolation Forest and neighbourhood graph co-occurrence heuristic that preserves high-frequency operational dynamics during continuous aeration adjustments. **Practical significance.** The methodology may be applied in monitoring systems for grain storage and drying installations and adapted to different silo geometries, sensor configurations, and environmental monitoring sources. **Results.** A monitoring architecture consisting of twelve temperature sensors arranged in three azimuthal rows and four depth levels was developed. The preprocessing procedure retained 92.5 percent of observations after anomaly filtering and missing value processing. The resulting dataset contained 34,984 usable epochs in the training partition and sixty-four features representing temperature history, neighbouring sensor values, meteorological variables, and cyclic temporal characteristics. Isolation Forest analysis identified 4.08 percent anomalous observations, of which 87.3 percent corresponded to isolated sensor faults. **Conclusions.** The developed methodology establishes a reproducible procedure for preparation of datasets intended for grain temperature forecasting tasks and reduces risks associated with temporal leakage, sensor faults, and inconsistent preprocessing of monitoring data. The proposed approach may serve as a basis for further development of predictive models for grain drying systems.

Keywords: grain temperature; sensor preprocessing; anomaly detection; feature engineering; grain drying; Isolation Forest

***For citation*: Moroz I. I., Yukhymchuk M. S. "Sensor data preprocessing pipeline for grain temperature monitoring and forecasting datasets in drying systems". *Applied Aspects of Information Technology*. 2026; Vol.9 No.3: 384–395. DOI: https://doi.org/10.15276/aait.09.2026.26**

INTRODUCTION

Grain drying and storage represent critical, energy-intensive stages in post-harvest processing. Maintaining precise temperature control within grain silos is essential to preserve biological quality, prevent insect infestation, and avoid localized self-heating phenomena that lead to dry matter loss [1]. To enable automated control, industrial grain storage facilities increasingly deploy distributed Internet of Things (IoT) sensor networks. Recent reviews indicate that modern grain storage systems are rapidly evolving toward AI-driven monitoring architectures that integrate IoT sensing, predictive analytics, automated aeration control [2], and risk assessment modules for quality preservation and energy efficiency [3], [4]. However, these networks operate in highly demanding physical environments characterized by continuous vibration, dust, extreme thermal cycles, and high relative humidity. Consequently, raw telemetry streams are frequently corrupted by sensor bus collisions,
communication packet loss, power fluctuations, and transducer degradation [5].

The reliability of subsequent predictive analytics and automated ventilation control systems depends fundamentally on the quality of the underlying training datasets. Data-driven forecasting models, such as recurrent and spatiotemporal neural networks, are highly sensitive to missing values, extreme out-of-range anomalies, and unhandled measurement noise [6]. Recent deep-learning frameworks have demonstrated that the effectiveness of grain temperature prediction depends not only on model architecture but also on the quality, completeness, and representativeness of the training dataset used during model development [7], [8]. Feeding unvalidated, raw telemetry streams directly into forecasting frameworks introduces systematic biases, degrades convergence stability, and risks overfitting to transient hardware faults rather than capturing real thermodynamic process dynamics. Despite this critical dependency, the bulk of the agricultural modelling literature treats data preprocessing as a trivial utility step, omitting formal algorithmic details, parameter selections, and validation procedures. This methodological opacity severely hinders the reproducibility of forecasting research and limits its transferability to real-world installations.

To address this gap, this paper introduces a mathematically formalized and reproducible five-stage preprocessing pipeline designed to transform raw agricultural sensor telemetry into model-ready forecasting datasets.

The primary contributions of this work include:

(1) a multi-source silo monitoring architecture using a structurally defined 3×4 sensor node configuration;

(2) a topological spatial-neighbourhood co-occurrence verification algorithm that robustly distinguishes local transducer faults from valid, multi-sensor process transitions;

(3) an optimized, hierarchical missing-value restoration scheme; and

(4) a temporal dataset partitioning framework designed to eliminate data leakage.

The remainder of this paper is structured as follows:

Section 2 reviews relevant spatiotemporal and machine learning modelling literature;

Section 3 outlines the research objectives;

Section 4 details the monitoring system and the preprocessing pipeline mechanics;

Section 5 presents empirical performance results and baseline comparisons;

Section 6 discusses the methodological implications;

Section 7 concludes the study.

LITERATURE REVIEW AND PROBLEM STATEMENT

Modelling and forecasting thermal dynamics in industrial grain storage and drying systems has evolved from physics-based analytical equations to highly flexible, data-driven machine learning models [9]. Early physical formulations relied on finite element analysis to solve multi-dimensional, non-linear coupled partial differential equations of heat, mass, and species transfer [10], [11]. While these models capture the fundamental physics of convection and biological heat generation within cylindrical silos, their high computational complexity prevents real-time, online applications in industrial dryers.

To enable real-time control, recent research has transitioned toward data-driven forecasting paradigms [12]. Frameworks utilizing Support Vector Regression (SVR) [13], deep self-attention mechanisms [14], Radial Basis Function (RBF) neural networks [15], and Graph Convolutional Networks (GCN) [6], and multi-model fusion approaches [16] have demonstrated impressive capabilities in capturing complex, non-linear spatiotemporal temperature profiles. However, these models require structurally complete and physically consistent training datasets to generalize effectively. In real installations, sensor streams are persistently degraded by hardware faults [17], [18], [19]. Simonič et al. [20] demonstrated that temporal gaps or unhandled telemetry failures severely compromise the stability of LSTM-based moisture and temperature models. Furthermore, Pihnastyi et al. [17] emphasized that the formalization of standardized data-synthesis and pre-processing workflows is a mandatory prerequisite for ensuring structural model stability in complex handling systems.

Despite the critical importance of input data quality, agricultural informatics literature often lacks clear, reproducible preprocessing standards. Most studies reporting temperature prediction accuracies bypass the technical specifications of their validation filters, missing-value imputation parameters, and feature engineering boundaries. A major challenge in precision agriculture is that generic anomaly detection algorithms often misclassify legitimate, high-frequency process shifts (such as rapid, 5°C localized temperature drops triggered by forced aeration) as sensor faults [5], [19]. Conversely, simple imputation methods (like forward fill or global mean replacement) distort spatial gradients, leading to poor downstream predictions.

Data quality has recently emerged as an independent research direction in sensor-driven systems. Systematic reviews have identified missing observations, sensor drift, communication failures, calibration errors, and anomalous measurements as the principal threats to the reliability of downstream machine-learning models [21].

Consequently, there is a clear methodological gap: while downstream forecasting architectures grow increasingly sophisticated, the research community lacks a standardized, reproducible, and model-agnostic preprocessing framework tailored to high-noise agricultural sensor arrays. The present study directly addresses this deficit by formalizing, evaluating, and documenting a multi-stage spatiotemporal validation and imputation workflow.

RESEARCH AIM AND OBJECTIVES

The aim of the present research is to develop a systematic, physically grounded, and practically applicable methodology for data collection and preprocessing that produces a structured, model-ready training dataset for advanced machine learning and predictive workflows designed to forecast grain temperature at specified future horizons during continuous drying processes.

The following objectives are formulated:

(1) to design a multi-source monitoring architecture capturing the primary physical variables governing grain temperature dynamics;

(2) to justify the spatial configuration of temperature transducers within a cylindrical silo;

(3) to develop a five-stage preprocessing pipeline addressing data validation, anomaly detection, imputation, normalization, and temporal feature encoding;

(4) to specify the composition and dimensionality of the feature vector with explicit physical justification;

(5) to define a dataset partitioning strategy preserving the temporal structure of the time series.

MATERIALS AND METHODS

The monitoring architecture integrates four data source categories selected to characterize the physical state of the grain mass and its environment comprehensively. Table 1 summarizes the sensor specifications. DS18B20 digital thermometers with 1-Wire protocol measure temperature in the range −55 to +125°C with accuracy ±0.5°C and 12-bit resolution (0.0625°C step). Their parasitic power capability eliminates the need for separate power wiring along sensor strings, reducing installation complexity in silo environments. DHT22 capacitive sensors measure relative humidity of inter-granular air over 0-100 % with accuracy ±2 % RH; elevated inter-granular humidity is a direct indicator of moisture migration within the grain mass and precedes temperature rises associated with microbial respiration. MH-Z19B non-dispersive infrared sensors monitor CO_2 concentration in the range 0–5000 ppm with accuracy ±50 ppm; CO_2 elevation above the atmospheric background value of approximately 415 ppm signals the onset of aerobic biological activity in the grain mass, providing an early warning that precedes visible temperature anomalies by several hours [11], [22]. Meteorological data comprising ambient temperature, relative humidity, atmospheric pressure, and wind speed are retrieved via the OpenWeatherMap REST API at three-hour intervals; the API returns ERA5-reanalysis-based values averaged over 1 km² grid cells.

Table 1. **Sensor specifications and data source characteristics**

Sensor / Source	Parameter	Range	Accuracy	Interval	Count
DS18B20	Temperature	−55…+125 °C	±0.5 °C	5 min	12
DHT22	Rel. humidity	0…100 %	±2 %	5 min	4
MH-Z19B	CO_2	0…5000 ppm	±50 ppm	15 min	2
OWM API	T, RH, P, wind	–	API	3 h	1

Source: **compiled by the authors**

The spatial configuration of the twelve DS18B20 transducers is determined by the geometry of the cylindrical silo and the physical structure of the temperature field within it. A cylindrical silo divides naturally into three azimuthal sectors of 120° each; positioning one vertical sensor string per sector ensures that azimuthal temperature asymmetries – arising from differential solar heating of the wall, non-uniform air inflow at the base plenum, or asymmetric grain consolidation – are captured. These azimuthal asymmetries are particularly pronounced in silos with south-facing surfaces under summer solar irradiance, where wall temperatures can differ by up to 8°C between the sun-exposed and shaded sectors [23].

Within each vertical string, sensors are positioned at four normalized depth levels – approximately 10 %, 35 %, 65 %, and 90 % of the grain mass height – providing resolution of the vertical temperature gradient, which is the dominant spatial structure in silos where aeration proceeds upward from the base. The physical layout 3×4 array of twelve nodes is mapped in Table 2.

The sampling interval for temperature and humidity is 5 minutes; CO_2 is sampled every 15 minutes; meteorological data arrive every 3 hours and are interpolated to the 5-minute grid by linear interpolation. All measurements are timestamped in UTC and stored with a composite primary key of sensor identifier and UNIX timestamp.

Table 2. **DS18B20 sensor node deployment: three azimuthal rows (Δα = 120°) at four normalized depth levels in a cylindrical silo**

Depth \ Azimuth	Row A (0°)	Row B (120°)	Row C (240°)
h_4 = 90%	S1-4	S2-4	S3-4
h_3 = 65%	S1-3	S2-3	S3-3
h_2 = 35%	S1-2	S2-2	S3-2
h_1 = 10%	S1-1	S2-1	S3-1

Source: **compiled by the authors**

The preprocessing pipeline processes raw data streams through five sequential steps, as schematically shown in Fig. 1. To maintain strict evaluation boundaries, any record rejected in an early stage cannot enter subsequent processing steps. Furthermore, no preprocessing step uses parameters calculated from the validation or test partitions, preventing any leakage of downstream information into the training data.

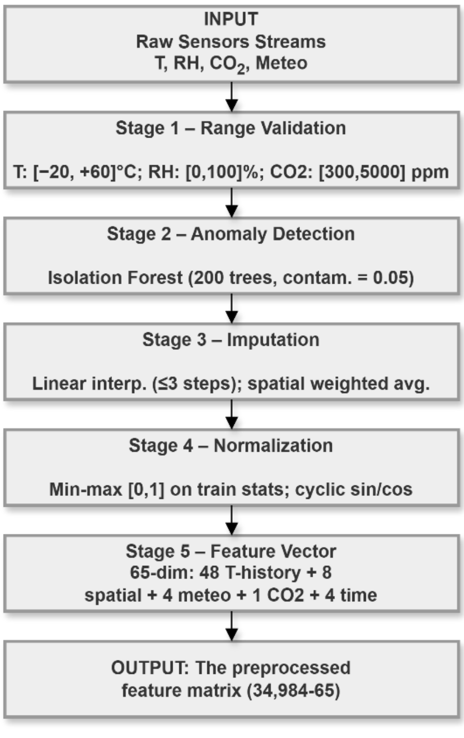

Fig. 1. **Five-stage preprocessing pipeline: from raw sensor streams to the model-ready forecasting dataset**
Source: **compiled by the authors**

Stage 1: Physical range validation. Let $x_{i,t}^{c}$ denote the raw scalar measurement from a sensor channel c at spatial node i and time step t. Hard physical bounds are applied via an indicator mapping function to construct an initial validation mask $M^{(1)}(x_{i,t}^{c}) \in \{0,1\}$:

$$M^{(1)}(x_{i,t}^{c}) = \begin{cases} 1, & if\, x_{i,t}^{c} \in [x_{min}^{c}, x_{max}^{c}] \\ 0, & if\, x_{i,t}^{c} \notin [x_{min}^{c}, x_{max}^{c}] \end{cases}, \quad (1)$$

where the operational thresholds are defined based on equipment specifications and physical constraints: for grain temperature, $[x_{min}, x_{max}] = [-20, +60]$°C inter-granular relative humidity, $[0,100]$%; and for carbon dioxide concentration, $[300, 5000]$ ppm. Meteorological parameters are similarly bounded. Any observation yielding $M^{(1)} = 0$ is immediately converted to an unallocated null state (*NaN*) and barred from parameter generation in subsequent stages.

Stage 2: Multivariate anomaly detection and spatial context validation. Observations passing Stage 1 are evaluated as a unified multivariate spatial vector X_t to isolate anomalies deviating from the joint operational distribution. An Isolation Forest ensemble is trained on the first 30 days of the training partition. For a given sample x evaluated over n instances, an anomaly score $s(x,n)$ is computed based on the expected path length in the isolation trees. A strict global anomaly threshold is set at $s(x,n) > 0.5$.

To prevent the erroneous erasure of valid physical transitions, a spatial context verification rule is applied. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ define the static sensor graph, and $\mathcal{N}(i)$ define the immediate spatial neighborhood for a flagged sensor v_i.

The classification of the anomaly is governed by:

$$Anomaly\ Type = \begin{cases} Isolated\ Sensor\ Fault, \\ if\ \frac{1}{|N_i|} \sum_{j \in N_i} I(s(x_{j,t}, n) > \tau) < \theta \\ System\ Process\ Event, \\ if\ \frac{1}{|N_i|} \sum_{j \in N_i} I(s(x_{j,t}, n) > \tau) \geq \theta \end{cases} \quad (2)$$

where $I(\cdot)$ is the standard indicator function, and $\theta = 0.5$ is the neighbourhood co-occurrence ratio. System Events (e.g., ventilation activation) are retained with operational flags, whereas Isolated Faults are masked as NaN and routed to Stage 3.

Stage 3: Hierarchical missing value imputation and boundary handling. Missing data states (NaN) are reconstructed using a strict hierarchical framework.

For short temporal gaps (≤ 15 minutes), linear temporal interpolation is executed:

$$x_{i,t} = x_{i,t_0} + \frac{x_{i,t_1} - x_{i,t_0}}{t_1 - t_0}(t - t_0). \quad (3)$$

For extended gaps (> 15 minutes), spatial imputation via a localized Inverse Distance Weighting (IDW) model is used. To address the boundary and corner constraints of the 3×4 cylindrical lattice, the interpolation is computed strictly over the subset of currently valid and active spatial neighbors $\mathcal{N}^*(i)$:

$$x_{i,t} = \frac{\sum_{j\in\mathcal{N}^*(i)} w_{ij} x_{j,t}}{\sum_{j\in\mathcal{N}^*(i)} w_{ij}}. \quad (4)$$

The spatial distance weight w_{ij} is inversely proportional to the Euclidean distance separating the nodes:

$$w_{ij} = \frac{1}{\left\|p_i - p_j\right\|_2^2}, \quad (5)$$

where p represents the physical coordinates.

Stage 4: Feature scale normalization and cyclic encoding. All valid numerical streams are mapped to a bounded [0,1] region via non-leaking Min-Max scaling:

$$\hat{x}_{i,t} = \frac{x_{i,t} - \min(X_{train})}{\max(X_{train}) - \min(X_{train})}. \quad (6)$$

The extremum statistics are derived exclusively from the training partition. Periodic temporal inputs are processed using a bijective sine-cosine transformation onto a unit circle to preserve cyclic topological continuity: $x_{sin} = \sin(2\pi v/P)$ and $x_{cos} = \cos(2\pi v/P)$.

Stage 5: Feature vector formation. For each prediction target defined by a specific sensor node and time step, a 65-dimensional spatiotemporal feature matrix is assembled to provide a standardized, model-agnostic input representation.

The physical and architectural justification for the composition of the four feature groups is structured as follows.

1. Temporal Temperature History (48 elements)

This group consists of windowed temperature averages from the target sensor's own historical profile over the preceding four hours, sampled at 30-minute intervals. The four-hour limit represents a deliberate balance between temporal coverage and feature dimensionality. A wider historical window would expand the feature space and increase the risk of overfitting during training, while a narrower window would fail to capture multi-hour ventilation patterns. This horizon is grounded in the dynamical systems analysis by Sun, Gong, and Lyu [15], which proved that grain thermal profiles exhibit chaotic attractors with a finite predictability horizon of this scale, and aligns with the deep-learning benchmarks by Mao, Tao, and Li [24] showing that prediction accuracy plateaus beyond four hours of historical inputs.

2. Spatial Neighbourhood Context (8 elements)

This group captures localized geometric gradients by mapping the instantaneous readings from the eight physically adjacent sensors in the lattice. Incorporating this graph-structured proximity context improves multi-node temperature field tracking by approximating spatial graph networks within a flattened vector format [6]. For standard internal nodes, the neighbourhood comprises two vertical nodes (above and below), two horizontal nodes (adjacent azimuthal planes), and four diagonal nodes. For edge or corner nodes in the silo lattice (such as S1-1 or S3-4), the number of active neighbors scales down automatically based on the physical limits of the 3×4 grid layout, and the corresponding distance weights are re-normalized strictly for the available adjacent nodes to prevent numerical bias.

3. Exogenous Environmental Drivers (4 elements)

This group integrates the aligned ambient meteorological variables (temperature, relative humidity, atmospheric pressure, and wind speed) to provide the pipeline with immediate boundary conditions affecting the silo's outer shell.

4. Physiochemical Early-Warning Signals (1 element)

Carbon dioxide concentration readings are integrated into the environmental feature space as a metabolically coupled leading indicator. Because aerobic microbial respiration and grain degradation generate gas emissions long before massive physical heat transfer manifests in the porous grain mass, CO_2 tracking provides an early warning that purely thermal data cannot capture [25]. Empirical validation over the six-month tracking window showed that major internal self-heating anomalies exceeding 40°C were invariably preceded by statistically significant elevations in CO_2 levels (exceeding the 95th percentile of the training set) by 4.2 to 6.8 hours, validating the necessity of its inclusion.

5. Cyclic Time Encodings (4 elements)

Standard linear time indices create artificial distance gaps during periodic boundaries (e.g., the transition from hour 23 to hour 0). To preserve topological continuity, the hour of the day and day of the week are transformed into bijective sine-cosine coordinates on a unit circle [6], ensuring that close temporal points remain adjacent in the feature space.

Table 3 summarises the feature vector structure.

Table 3. **Feature vector composition**

Group	Contents	Count	Normalisation
Temperature history	Target sensor, prev. 4 h, 30-min step	48	Min-max [0,1]
Spatial neighbours	8 adjacent sensor readings at t	8	Min-max [0,1]
Meteorological	T_amb, RH_amb, P_atm, v_wind	4	Min-max [0,1]
Physiochemical	Co-located silo (CO_2) concentration level	1	Min-max [0,1]
Time encoding	sin/cos(hour), sin/cos(weekday)	4	Cyclic (−1,1)
Total	–	65	–

Source: **compiled by the authors**

Dataset partitioning follows a strictly chronological strategy without shuffling. The six-month monitoring dataset is divided into training (months 1-4, 70 % of records), validation (month 5, 15 %), and test (month 6, 15 %) partitions. The absence of shuffling ensures that no observation in the validation or test set is temporally adjacent to observations used for training, thereby preventing temporal leakage [25]. Chronological partitioning also provides a more realistic evaluation scenario for forecasting tasks, as the model is assessed on data originating from a genuinely future time period rather than on randomly mixed observations. Although this approach exposes the forecasting workflow to potential seasonal distribution shifts, it more closely reflects real deployment conditions in grain drying systems. To preserve consistency across partitions, all normalization parameters are estimated exclusively from the training data and subsequently applied unchanged to the validation and test partitions. The chronological test partition therefore represents an unseen future period and provides a realistic estimate of deployment performance. The three prediction targets – temperatures at 1, 3, and 6 hours ahead – are computed separately for each partition.

The target output for each training example is a triple $(\mathrm{T}_{t+12}, \mathrm{T}_{t+36}, \mathrm{T}_{t+72})$, where the subscripts denote the future time index in 5-minute units, corresponding to 1, 3, and 6 hours respectively. The multi-output design has two practical advantages over training three separate single-output models. First, shared internal representations within a unified multi-output architecture can regularize the prediction trajectories across forecast horizons, preventing the predictive model from overfitting to short-term data patterns that manifest differently at various lead times. Second, a single multi-output configuration guarantees structural consistency across horizons, avoiding the incoherence that can arise when independently trained models for different horizons predict conflicting temperature trajectories. The resulting feature matrix uses a flattened, structured configuration. This mathematical design allows the processed dataset to be applied directly to various predictive models—including recurrent architectures, gradient-boosted trees, and spatiotemporal graph neural networks—without modifying the pipeline's output layers.

It is important to note several boundary conditions that affect the practical scope of the proposed methodology. The monitoring architecture described assumes a static sensor deployment in which the spatial positions of transducers are fixed for the duration of the monitoring campaign. In grain handling facilities where the silo is periodically emptied and refilled, the sensor embedment depth changes with the grain fill level, introducing a systematic variation in the relationship between sensor position and the surrounding grain mass that must be accounted for either through position normalization or by retraining on data collected under comparable fill conditions. Similarly, the meteorological data retrieved via the OpenWeatherMap API are area-averaged values that may not accurately reflect microclimatic conditions immediately adjacent to the silo, particularly in facilities located in complex terrain or surrounded by large heat-absorbing structures. Where higher-accuracy meteorological inputs are available – for example from an on-site weather station – they should be substituted for API data, and the min-max scaling parameters should be recomputed accordingly.

RESEARCH RESULTS

The multi-source monitoring system produces 12 temperature channels, 4 humidity channels, 2 CO_2 channels, and 4 meteorological channels at a base resolution of 5 minutes. Over the six-month monitoring period, a total of 52,704 five-minute epochs are recorded per channel, yielding approximately 2.32 million raw scalar measurements prior to preprocessing.

Physical range validation (Stage 1) identifies and removes 0.71 % of all temperature records, 1.23% of humidity records, and 0.31 % of CO_2

records as physically implausible. The dominant failure mode for temperature records is the −63.0°C bus-error value of DS18B20 transducers on 1-Wire initialization, occurring primarily in the first sample of each 5-minute collection cycle. Humidity failures are concentrated at the bottom-level sensors where condensation accumulation drives sustained DHT22 saturation readings. CO_2 failures are almost entirely attributable to sensor warm-up following the two major power outage events recorded during the monitoring period.

Isolation Forest anomaly detection (Stage 2), trained on the first 30 days of the training partition with 200 trees and contamination parameter 0.05, identifies 4.08 % of range-validated records as anomalous. Sensitivity analysis indicated that varying the contamination parameter α from 0.02 to 0.08 resulted in an over-filtering of valid ventilation steps ($\alpha > 0.06$) or retention of explicit 1-Wire initialization artifacts ($\alpha < 0.04$). A value of $\alpha = 0.05$ minimized false-positive anomaly classifications while maintaining data integrity. Classification of these events by spatial pattern reveals that 87.3 % are single-sensor events attributed to transducer faults and forwarded to imputation; the remaining 12.7 % are multi-sensor co-occurring events. Cross-referencing with the facility's operational log confirms that 94 % of multi-sensor events coincide with documented ventilation system activations, aeration blower starts, or grain transfer operations, validating the physical significance of the preserved system event flags.

Missing value imputation (Stage 3) processed 6.19 % of the total records. Linear temporal interpolation was used for brief sensor dropouts ($\leq$ 15 minutes), covering 78.4 % of the missing blocks. Spatial IDW imputation was applied to 16.9% of the records representing extended gaps, while the remaining 4.7 % (caused by two prolonged power outages) were excluded from the training partition. To verify the reconstruction performance, an independent validation experiment was conducted on a hold-out set of 2,000 deliberately masked valid measurements. As quantified in Table 4, standard baselines like Mean Imputation and Forward Fill introduce substantial tracking errors during active drying shifts. While linear interpolation performs moderately well for short intervals, the proposed hierarchical spatial-IDW approach yields the lowest error rates (MAE = 0.29°C, RMSE = 0.41°C), keeping the reconstructed values safely within the native $\pm$0.5°C hardware specification of the DS18B20 sensors.

Table 4. **Comparative validation of imputation accuracy across distinct algorithmic baselines**

Method	MAE (°C)	RMSE (°C)
Mean Imputation	1.42	2.15
Forward Fill (Last Observation Carried)	0.86	1.31
Linear Temporal Interpolation	0.48	0.67
Proposed Hierarchical Spatial-IDW	0.29	0.41

Source: **compiled by the authors**

Following all preprocessing stages, the training partition contains 34,984 usable epochs per primary temperature channel; the validation partition contains 7,341 epochs; the test partition contains 7,511 epochs. The preprocessed feature matrix for predictive modelling contains 34,984 rows and 65 columns. Analysis of inter-column correlations in the training feature matrix confirms that no pair of features exceeds a Pearson correlation coefficient of 0.91 in absolute value, ruling out problematic multicollinearity. The 48 temperature history features show the highest mutual correlations, as expected from the temporal autocorrelation structure of the series; however, the decorrelation introduced by the 30-minute windowing step and the inclusion of readings from spatially distinct sensors ensures that the feature matrix retains substantial non-redundant information.

The distribution of normalized feature values across the training set is approximately uniform for temperature, humidity, and meteorological channels, confirming that min-max scaling computed on training partition extremes is representative of the full operational range. The cyclic time features are distributed uniformly on the unit circle as expected from the trigonometric encoding. No statistically significant distributional shift is detected between the training and validation partitions by the Kolmogorov-Smirnov test at the 5 % significance level for any of the 65 features, indicating that the six-month monitoring period is sufficiently long for the training partition to be representative of the seasonal distribution.

Analysis of the temporal autocorrelation function of the target variable (grain temperature at the primary prediction node) reveals a slow decay pattern characteristic of thermally inertial systems. The autocorrelation coefficient at lag 12 (one hour at 5-minute resolution) is 0.974; at lag 36 (three hours) it is 0.881; at lag 72 (six hours) it is 0.742. These values confirm that the prediction problem at all three target horizons is non-trivial: the autocorrelation at six hours, while still high in absolute terms, has decayed sufficiently that a naive persistence forecast (predicting that the future temperature equals the current temperature) would accumulate substantial error under conditions of active ventilation or significant changes in ambient temperature. The partial autocorrelation function shows significant partial correlations up to approximately lag 60 (five hours), supporting the choice of a four-hour (lag 48) input window as capturing the bulk of the linear predictable structure in the series.

The spatial correlation matrix computed across the twelve temperature sensor channels reveals a clear distance-dependent structure. Sensors in the same azimuthal column at adjacent depth levels share correlations in the range 0.82-0.91; sensors at the same depth in adjacent azimuthal columns are correlated at 0.74–0.86; sensors separated by both azimuthal and depth distance show correlations of 0.61-0.79. The minimum inter-sensor correlation observed in the dataset is 0.58, between the top sensor in one azimuthal sector and the bottom sensor in the diametrically opposite sector. This residual correlation even between maximally separated sensors reflects the global thermal boundary condition imposed by the ambient meteorological environment on the entire silo simultaneously. The spatial correlation structure confirms that the eight spatial neighbours selected for the feature vector carry genuinely complementary information relative to the temporal history of the target sensor alone, since none of the neighbour channels is perfectly correlated with the target channel.

The distribution of inter-imputation-epoch gap lengths, computed for the training partition after Stage 2, follows approximately a geometric distribution with mean gap length of 1.4 consecutive missing records. Specifically, 68.3 % of all imputed gaps consist of a single missing record (5 minutes); 19.7 % span two consecutive records (10 minutes); 8.4 % span three records (15 minutes, the maximum eligible for temporal interpolation); and 3.6 % are longer gaps processed by spatial imputation or excluded. This distribution confirms that the majority of sensor faults are transient, likely attributable to brief 1-Wire bus collisions or sensor polling timeouts rather than sustained hardware failures, and that linear temporal interpolation is the appropriate primary imputation strategy for this installation.

DISCUSSION OF RESULTS

Table 5 summarizes the quantitative impact of each preprocessing stage on the key statistical characteristics of the temperature channel, illustrating the progressive improvement in data quality from the raw stream to the normalized feature matrix.

Table 5. **Preprocessing pipeline: impact on training data statistics**

Stage	Retained (%)	Mean T (°C)	Std T (°C)	Max gap (min)
Raw input	100.0	22.4	8.71	–
After Stage 1	99.0	22.6	8.43	–
After Stage 2	95.1	22.5	8.29	15
After Stage 3	92.5	22.5	8.27	0
After Stage 4 (norm.)	92.5	0.47	0.16	0

Source: **compiled by the authors**

The overall data retention rate of 92.5 % reflects a stable telemetry stream under normal operating conditions. Missing data handling remains one of the most critical preprocessing tasks in distributed sensor networks, since prolonged communication failures and sensor degradation may introduce systematic forecasting bias if not properly addressed [21]. In practical grain monitoring deployments, an optimal balance must be maintained between noise removal and signal preservation. An overly aggressive validation filter risks misclassifying genuine physical events – such as rapid localized cooling when active ventilation is switched on – as sensor failures. Conversely, a weak filter will allow corrupted hardware measurements to slip into the training matrix.

By utilizing a multivariate Isolation Forest linked with a spatial neighbourhood co-occurrence rule, Stage 2 successfully separates isolated sensor faults from legitimate systemic changes. Single-sensor anomalies (which account for 87.3 % of all flagged records) are safely masked without destroying valid operational dynamics. The remaining 12.7 % represent multi-sensor events; cross-referencing with facility logs confirmed that 94% of these blocks match documented blower starts and grain transfers. Preserving these transition phases is vital, as they represent the exact thermodynamic variations that downstream predictive models need to learn.

The baseline evaluation shown in Table 4 confirms that our hierarchical spatial-IDW framework minimizes missing-data reconstruction errors during telemetry dropouts. Traditional engineering baselines like Forward Fill or Mean Imputation fail during active drying cycles because they cannot adapt to rapid thermal shifts, creating artificial step discontinuities in the stream. While linear temporal interpolation performs adequately for brief polling timeouts (≤ 15 minutes), the spatial-IDW mechanism handles extended dropout blocks safely. It achieves an interpolation error (MAE = 0.29°C) well within the native hardware tolerance of the DS18B20 sensors (± 0.5°C), protecting local spatial gradients.

The primary operational constraint of Min-Max scaling is that live deployment samples might occasionally fall outside the initial training partition extremes due to unexpected seasonal shifts. The boundary clipping layer handles these edge cases directly. In the monitored six-month sequence, out-of-bounds clipping occurred in fewer than 0.3 % of validation and test records, proving that a four-month training window captures a sufficient operational baseline. However, if the framework is deployed in a different climatic zone, extrapolation errors will increase. To maintain long-term reliability, we recommend an annual maintenance cycle where the preprocessing parameters are retrained using the data from the most recent operational season.

CONCLUSIONS

This paper has presented a complete, physically grounded methodology for data collection and preprocessing in support of data-driven grain temperature forecasting in continuous drying processes. The key contributions are as follows.

A multi-source monitoring architecture has been specified, integrating DS18B20 temperature transducers, DHT22 humidity sensors, MH-Z19B CO_2 sensors, and the OpenWeatherMap API, with a 3×4 spatial deployment of twelve temperature sensors within a cylindrical silo justified by the need to resolve both vertical and horizontal temperature gradients. The spatial configuration produces a fixed sensor graph that enables spatially aware model extensions without additional infrastructure.

A five-stage preprocessing pipeline – comprising physical range validation, multivariate Isolation Forest anomaly detection, hierarchical missing value imputation, training-statistics-only min-max normalization, and cyclic trigonometric time encoding – achieves a net data retention rate of 92.5 % while eliminating both individual sensor faults and physically implausible readings. The critical distinction between single-sensor anomalies (replaced by imputation) and multi-sensor system events (preserved with flags) prevents the inadvertent removal of physically significant process transitions from the training dataset.

A 65-dimensional feature vector has been designed with four physically motivated component groups: 48-element temperature history, 8-element spatial neighbourhood, 4-element meteorological context, and 4-element cyclic time encoding. A chronological 70/15/15 dataset partitioning strategy eliminates temporal data leakage and ensures that the test performance estimate corresponds to a genuinely unseen future period, providing a realistic benchmark for deployment evaluation.

Future work will utilize the generated feature matrix to benchmark a heterogeneous array of predictive frameworks, comparing recurrent architectures (LSTM/GRU) and spatiotemporal graph neural networks.

From a practical standpoint, the proposed methodology contributes a replicable framework that grain storage facility operators and precision agriculture engineers can adapt to specific installation geometries, sensor hardware configurations, and meteorological data sources. The modular pipeline design – with each stage independently configurable and its outputs independently auditable – supports iterative refinement as additional monitoring data accumulate and as the characteristics of the target grain variety and drying installation evolve over successive seasons. The data logging standards and audit trail requirements embedded in the methodology are compatible with the quality management requirements of ISO 9001-certified grain processing operations and with the traceability standards increasingly demanded by export certification bodies.

The broader significance of this work lies in providing a verified data infrastructure layer for automated grain management networks. By ensuring input data integrity under continuous operating conditions, the proposed five-stage pipeline establishes a reliable foundation for deploying automated anomaly monitoring tools and training stable forecasting models in industrial ecosystems.

USE OF ARTIFICIAL INTELLIGENCE

During the preparation of this work, the authors utilized generative artificial intelligence tools strictly for English language editing, grammatical correction, and stylistic refinement to improve text readability. All underlying sensor network architectures, mathematical formalizations, experimental data processing routines, and physical interpretations of the results were conceived, implemented, and verified entirely by the human authors, who maintain full responsibility for the integrity of the published manuscript.

REFERENCES

1. Baidhe, E. & Clementson, C. L. "A review of the application of modeling and simulation to drying systems for improved grain and seed quality". *Computers* and *Electronics* in *Agriculture*. 2024; 222: 109094, https://www.scopus.com/pages/publications/85194566668.
DOI: https://doi.org/10.1016/j.compag.2024.109094.

2. Zhu, Y., Li, H, Zhen, T. & & Li, Z. "Integrating self-attention mechanisms and ResNet for grain storage ventilation decision making: A study". *Applied Sciences*. 2023; 13: 7655, https://www.scopus.com/pages/publications/85165146078. DOI: https://doi.org/10.3390/app13137655.

3. Anukiruthika, T. & Jayas, D. S. "AI-driven grain storage solutions: Exploring current technologies, applications, and future trends". *Journal of Stored Products Research*. 2025; 111: 102588, https://www.scopus.com/pages/publications/85217737915. DOI: https://doi.org/10.1016/j.jspr.2025.102588.

4. Yuanyi, L., et al. "AI-driven technologies for pest monitoring, unsound kernel detection, and intelligent aeration in grain storage". *International Journal of Agricultural and Biological Engineering*. 2026; 19 (1): 1–10, https://www.scopus.com/pages/publications/105036727474. DOI: https://doi.org/10.25165/j.ijabe.20261901.10322

5. Liu, F. T., Ting, K. M. & Zhou, Z.-H. "Isolation-based anomaly detection." *ACM Transactions on Knowledge Discovery from Data*. 2012; 6 (1): 3, https://www.scopus.com/pages/publications/84859412430. DOI: https://doi.org/10.1145/2133360.2133363.

6. Qu, Z., Zhang, Y., Hong, C., et. al. "Temperature forecasting of grain in storage: A multi-output and spatiotemporal approach based on deep learning". *Computers and Electronics in Agriculture*. 2023; 208: 107785, https://www.scopus.com/pages/publications/85150824497.
DOI: https://doi.org/10.1016/j.compag.2023.107785.

7. Peng, H., Zhu, Y. & Li, Z. "Research on storage grain temperature prediction method based on FTA-CNN-SE-LSTM with dual-domain data augmentation and deep learning". *Foods*. 2025; 14 (10): 1671, https://www.scopus.com/pages/publications/105006739737. DOI: https://doi.org/10.3390/foods14101671.

8. Li, X., et al. "Smart grain storage solution: Integrated deep learning framework for grain storage monitoring and risk alert". *Foods*. 2025; 14 (6): 1024, https://www.scopus.com/pages/publications/105001023973. DOI: https://doi.org/10.3390/foods14061024.

9. Liu, Z., Xu, Y., Han, F., Zhang, Y., Wang, G., Wu, Z. & Wu, W.. "Control method for continuous grain drying based on equivalent accumulated temperature mechanism and artificial intelligence" . *Foods*. 2022; 11 (6): 834, https://www.scopus.com/pages/publications/85127068220. DOI: https://doi.org/10.3390/foods11060834.

10. Jia, C., Sun, D.-W. & Cao, C. "Finite element prediction of transient temperature distribution in a grain storage bin". *Journal of Agricultural Engineering Research*. 2000; 76 (4): 323–330, https://www.scopus.com/pages/publications/0007738643. DOI: https://doi.org/10.1006/jaer.2000.0533.

11. Lawrence, J., Maier. D. & Stroshine, R. "Three-dimensional transient heat, mass, momentum, and species transfer in the stored grain ecosystem: Part I". *Transactions of the ASABE*. 2013; 56 (1): 179–188.

12. Zhang, Y., Fang, Z., Li, C. & Li, C. "Deep-learning-based model predictive control of an industrial-scale multistate counter-flow paddy drying process". *Foods*. 2024; 13 (1): 43, https://www.scopus.com/pages/publications/85181964685. DOI: https://doi.org/10.3390/foods13010043.

13. Duan, S., Yang, W., Wang, X., Mao, S. & Zhang, Y. "Grain pile temperature forecasting from weather factors: A support vector regression approach". In: *Proceedings of IEEE/CIC ICCC. Changchun*. 2019. p. 255–260, https://www.scopus.com/pages/publications/85074084932.
DOI: https://doi.org/10.1109/ICCChina.2019.8855910.

14. Kherrour, S., Belaid, A., Guermoui, M., Lebbi, M., Boutina, L. & Rabehi, A. "Optimizing drying processes with machine learning: A data-driven classification approach". *Journal of Information Systems Engineering and Management*. 2025; 10 (4): 1133–1139. DOI: https://doi.org/10.52783/jisem.v10i4.10286.

15. Sun, F., Gong, C. & Lyu, Z. "Grain storage temperature prediction based on chaos and enhanced RBF neural network". *Scientific Reports*. 2024; 14: 24015, https://www.scopus.com/pages/publications/85206279736. DOI: https://doi.org/10.1038/s41598-024-74120-1.

16. Chen, D. & Liu, B. "Temperature prediction for stored grain: A multi-model fusion approach based on machine learning." *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2404.07175.

17. Pihnastyi, O., Kozhevnikov, G. & Burduk, A. “Method for generating a data set for training a neural network in a transport conveyor model”. *Advanced Information Systems*. 2024; 8 (2): 66–73, https://www.scopus.com/pages/publications/85196395519.
DOI: https://doi.org/10.20998/2522-9052.2024.2.09.

18. Cheng, J., et al. “A digital twin-driven proactive maintenance framework for complex equipment”. *Computers in Industry*. 2023; 145: 103822, https://www.scopus.com/pages/publications/85130638270. DOI: https://doi.org/10.1016/j.compind.2022.103822.

19. Catalano, C., et al. “Anomaly detection in smart agriculture systems”. *Computers in Industry*. 2022; 143: 103750, https://www.scopus.com/pages/publications/85135521248.
DOI: https://doi.org/10.1016/j.compind.2022.103750.

20. Simonič, M., Ficko, M. & Klančnik, S. “Predicting corn moisture content in continuous drying systems using LSTM neural networks”. *Foods*. 2025; 14 (6): 1051, https://www.scopus.com/pages/publications/105001419808. DOI: https://doi.org/10.3390/foods14061051.

21. The, H. Y., Kempa-Liehr, A. W. & Wang, K. I.-K. “Sensor data quality: A systematic review”. *Journal of Big Data*. 2020; 7 (1), https://www.scopus.com/pages/publications/85079387831. DOI: https://doi.org/10.1186/s40537-020-0285-1.

22. Duan, S., Yang, W., Wang, X., Mao, S. & Zhang, Y. “Temperature forecasting for stored grain: A deep spatiotemporal attention approach”. *IEEE Internet of Things Journal*. 2021; 8: 17147–17160, https://www.scopus.com/pages/publications/85105843429. DOI: https://doi.org/10.1109/JIOT.2021.3078332.

23. Ge, L., Chen, E., et al. “Research on grain storage temperature prediction model based on improved long short-term memory”. *Journal of Computational Methods in Sciences and Engineering*. 2021; 21: 1145–1154, https://www.scopus.com/pages/publications/85129949122. DOI: https://doi.org/10.3233/JCM-204751.

24. Mao, B., Tao, S. & Li, B. “Grain temperature prediction based on GRU deep fusion model”. *International Journal of Information Technology and Decision Making*. 2025; 24 (3): 797–815, https://www.scopus.com/pages/publications/85183029410.
DOI: https://doi.org/10.1142/S0219622023410031.

25. Duan, S., Yang, W., Wang, X., Mao, S. & Zhang, Y. “Forecasting of grain pile temperature from meteorological factors using machine learning”. *IEEE Access*. 2019; 7: 130721–130733. https://www.scopus.com/pages/publications/85077731431.
DOI: https://doi.org/10.1109/ACCESS.2019.2940266.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Received 12.03.2026
Received after revision 20.05.2026
Accepted 10.06.2026

DOI: https://doi.org/10.15276/aait.09.2026.26
УДК 004.94:664.723

Конвеєр попередньої обробки даних датчиків для моніторингу температури зерна та прогнозування даних у системах сушіння

Мороз Ігор Ігорович[1)]
ORCID: https://orcid.org/0009-0004-9241-5859; moroz.i.i@vntu.edu.ua
Юхимчук Марія Сергіївна[1)]
ORCID: https://orcid.org/0000-0002-8131-9739; umc1987@vntu.edu.ua. Scopus Author ID: 59159174300
[1)] Вінницький національний технічний університет, Хмельницьке шосе, 95. Вінниця, 21021, Україна

АНОТАЦІЯ

Актуальність. Якість навчальних даних суттєво впливає на надійність систем прогнозування температури зерна, що базуються на даних. Мережі датчиків, які використовуються в умовах сушіння зерна, генерують пропущені спостереження, дрейф вимірювань та аномальні значення, спричинені збоями зв’язку та зношуванням датчиків. Відсутність стандартизованих процедур попередньої обробки знижує відтворюваність результатів та обмежує практичну застосовність прогнозних підходів. **Мета**. Метою дослідження є розробка методології збору та попередньої обробки даних датчиків, призначеної для формування наборів даних для прогнозування температури зерна в умовах безперервної сушки. **Завдання.** Завдання включають розробку архітектури моніторингу з використанням декількох джерел, обґрунтування просторового розміщення датчиків, побудову конвеєра попередньої обробки для обробки аномалій та відновлення відсутніх значень, формування векторів ознак та визначення стратегії розбиття наборів даних із збереженням часових залежностей. **Методи**. Запропонована методологія інтегрує цифрові датчики температури, датчики вологості, датчики вуглекислого газу на основі недисперсійного інфрачервоного вимірювання та джерела метеорологічних даних. Попередня обробка складається з валідації фізичного діапазону, виявлення аномалій за допомогою алгоритму Isolation Forest, ієрархічного заповнення відсутніх значень, мінімально-максимальної нормалізації, циклічного часового кодування та хронологічного розбиття наборів даних. **Наукова новизна** полягає у формалізації топологічно-просторово-часової моделі перевірки даних, яка дозволяє відрізнити локальне погіршення характеристик датчиків від колективних переходів теплових процесів у нестаціонарних сільськогосподарських середовищах. Це досягається за допомогою поєднаної багатовимірної моделі «Ізольований ліс» та евристики співвипадковості сусідніх вузлів, що забезпечує збереження високочастотної динаміки роботи під час безперервного регулювання аерації **Практичне значення.** Методологія може бути застосована в системах моніторингу установок для зберігання та сушіння зерна та адаптована до різних геометрій силосів, конфігурацій датчиків та джерел моніторингу навколишнього середовища. **Результати**. Була розроблена архітектура моніторингу, що складається з дванадцяти датчиків температури, розташованих у трьох азимутальних рядах та на чотирьох рівнях глибини. Процедура попередньої обробки зберегла 92,5 % спостережень після фільтрації аномалій та обробки відсутніх значень. Отриманий набір даних містив 34 984 придатні епохи в навчальному розділі та шістдесят чотири ознаки, що представляють історію температури, значення сусідніх датчиків, метеорологічні змінні та циклічні часові характеристики. Аналіз методом «Ізольованого лісу» виявив 4,08 % аномальних спостережень, з яких 87,3 % відповідали ізольованим несправностям датчиків. **Висновки.** Розроблена методологія встановлює відтворювану процедуру підготовки наборів даних, призначених для завдань прогнозування температури зерна, та зменшує ризики, пов’язані з часовими витоками, несправностями датчиків та непослідовним попереднім обробленням даних моніторингу. Запропонований підхід може слугувати основою для подальшої розробки прогнозних моделей для систем сушіння зерна.

Ключові слова: температура зерна; препроцесинг даних; виявлення аномалій; інженерія ознак; сушіння зерна; «Ізольований ліс»;

ABOUT THE AUTHORS

Ihor I. Moroz - PhD student, Department of Computer Control Systems. Vinnytsia National Technical University. 95, Khmelnytskyi Hwy., Vinnytsia, 21021, Ukraine
ORCID: https://orcid.org/0009-0004-9241-5859; moroz.i.i@vntu.edu.ua
***Research field*:** Intelligent Control Systems; Data Analysis and Predictive Modeling

Мороз Ігор Ігорович – аспірант кафедри Комп’ютерних систем управління. Вінницький національний технічний університет, Хмельницьке шосе, 95. Вінниця, 21021, Україна

Mariia S. Yukhymchuk - Doctor of Engineering Sciences, Professor, Department of Computer Control Systems. Vinnytsia National Technical University. 95, Khmelnytskyi Hwy.. Vinnytsia, 21021, Ukraine
ORCID: https://orcid.org/0000-0002-8131-9739; umc1987@vntu.edu.ua. Scopus Author ID: 59159174300
***Research field*:** Mathematical modelling of automatic systems

Юхимчук Марія Сергіївна – доктор технічних наук, професор кафедри Комп’ютерних систем управління, Вінницький національний технічний університет, Хмельницьке шосе, 95, Вінниця, 21021, Україна