Baltic J. Modern Computing, Vol. 14 (2026), No. 3, 669-683
https://doi.org/10.22364/bjmc.2026.14.3.08

Using Big Data Technologies for Intelligent Analysis and Real-Time Moisture Estimation in the Sugar Drying Process

Halyna GRYHORCHUK, Lyubomyr GRYHORCHUK,
Ruslan KHRABATYN, Mykhailo KRYKHIVSKYI

Ivano-Frankivsk National Technical University of Oil and Gas, Ukraine 76019, Ivano-Frankivsk, 15 Karpatska Street

halyna.hryhorchuk@nung.edu.ua, l.gryhorchuk@nung.edu.ua,
r.khrabatyn@nung.edu.ua, mykhailo.krykhivskyi@nung.edu.ua

ORCID 0000-0003-1674-9828, ORCID 0000-0003-0924-5090, ORCID 0000-0002-4418-1152, ORCID 0009-0000-3285-4308

Abstract. Industrial drying processes generate large volumes of time-series data via SCADA/PLC systems, yet this data is predominantly used for monitoring rather than intelligent analysis. This paper presents a process-oriented approach to sugar drying analysis and real-time moisture estimation, based on Big Data analytics and physics-guided machine learning. Archival SCADA data are preprocessed and transformed into structured datasets with physically justified features, including the temperature differential ΔT. A hybrid soft sensor combining a physical drying model with neural network correction is proposed for real-time moisture estimation. The model achieves MAE = 0.062%, RMSE = 0.089%, $R^2 = 0.97$, outperforming PID control, linear regression, and standalone ANN approaches. The principal contributions of this work are: (i) physics-guided feature engineering from archival SCADA data, including the temperature differential ΔT; (ii) k-means clustering for identification of operating regimes; (iii) a hybrid soft sensor combining a physics-based drying model with an MLP correction term for real-time moisture estimation; and (iv) an ablation study quantifying the contribution of the physical baseline and time-lagged features to overall accuracy.

Keywords: Big Data; SCADA; soft sensor; machine learning; hybrid modeling; physics-guided modeling; drying process.

1. Introduction

Drying processes are widely used in the food and chemical industries and are characterized by complex nonlinear relationships among heat carrier parameters, equipment operating conditions, and raw material properties. In sugar production, the final moisture content of the product is a key quality indicator that determines its storage stability, transportability, and energy consumption.

In industrial settings, direct measurement of product moisture is often difficult or is performed with a delay due to laboratory testing. For this reason, control systems rely on indirect parameters such as drying agent temperature, air flow rate, and process duration, which reduces the precision of regulation.

Modern automation systems (SCADA/PLC) ensure continuous accumulation of large volumes of process data in the form of time series (Wang et al., 2015). Such data contain information on temperature profiles, heat carrier flow rates, drum rotation speed, load, and other process parameters. These data arrays can be regarded as industrial Big Data (Kusiak, 2017).

Despite the substantial volume of available information, it is in most cases used only for monitoring and fault analysis (Fernández-Caramés et al., 2021). The application of data analysis methods, machine learning, and artificial intelligence opens new opportunities for identifying hidden patterns and optimizing technological processes. At the same time, purely data-driven approaches often fail to account for the physical nature of heat and mass transfer processes, which may lead to unstable or incorrect results. This motivates the theory-guided data science paradigm (Karpatne et al., 2017), in which domain knowledge is systematically integrated with data-driven discovery. Therefore, combining Big Data analysis with physically justified process features is a relevant challenge.

The objective of this work is to develop an approach to the analysis of the sugar drying process based on Big Data technologies that enables the extraction of informative features, identification of efficient operating regimes, and preparation of data for subsequent intelligent modeling and control. The main contributions of this work are as follows: (i) physics-guided feature engineering from archival SCADA data, including the temperature differential ΔT; (ii) k-means clustering for identification of operating regimes; (iii) a hybrid soft sensor combining a physics-based drying model with an MLP correction term for real-time moisture estimation; and (iv) an ablation study quantifying the contribution of the physical baseline and time-lagged features to overall accuracy.

2. Review of current approaches to drying process analysis

Current research on drying processes in the food and chemical industries focuses on improving energy efficiency, control accuracy, and final product quality. Traditional approaches are based on physical heat and mass transfer models that describe the drying process by means of differential equations (Mujumdar, 2015; Kudra and Mujumdar, 2009).

Classical drying models account for moisture evaporation kinetics, heat fluxes, and diffusion processes; however, their application under real industrial conditions is limited by the complexity of the process and the presence of uncertainties.

Recent studies have given considerable attention to the use of intelligent data analysis and machine learning methods. In particular, regression models, artificial neural networks, and ensemble methods have been applied to predict drying process parameters (Khan et al., 2022a; Kadlec et al., 2009; Khan et al., 2022b).

A separate line of research is concerned with the use of Big Data technologies for processing industrial data obtained from SCADA/PLC systems. Such approaches enable the identification of hidden patterns, optimization of equipment operating regimes, and improvement of prediction accuracy (Qin, 2012).

Recent studies increasingly employ hybrid approaches that combine physical models with machine learning methods (Miraei Ashtiani and Martynenko, 2024; Khan et al., 2022a; von Stosch et al., 2014). The mathematical modelling of the massecuite drying process in rotary drum dryers has been addressed in Gryhorchuk et al. (2025), providing the physical basis for the present work.

Despite significant progress, the challenge of aligning data-driven models with the physical nature of processes remains. This underlines the relevance of developing approaches that combine Big Data analysis with physically justified features of the drying process.

3. Intelligent data analysis and artificial intelligence

From an artificial intelligence perspective, archival SCADA data can be transformed into structured datasets suitable for the application of intelligent analysis methods.

The proposed approach involves the following methods: clustering of operating regimes based on temperature profiles and process parameters; construction of regression models for estimating process parameters; application of neural networks for modeling nonlinear relationships; detection of anomalies in equipment operation; analysis of the significance of process parameters.

To structure the archival data and identify characteristic operating regimes of the dryer, the k-means clustering algorithm was applied. The algorithm enabled the automatic identification of four technological process states — intensive drying, stabilisation, final stage, and cooling — which corresponds to the physical nature of the process and confirms the validity of the proposed model structure. The clustering results were used to form the training dataset and to verify the boundaries between drying stages.

An important element is the use of physically justified features such as temperature differentials, heat fluxes, and integral process characteristics, which ensures that the results of the analysis are consistent with the physical nature of the drying process (Khan et al., 2022a). Among the considered approaches, neural networks are of particular interest due to their ability to model nonlinear relationships.

Thus, Big Data technologies serve as the foundation for developing intelligent models and process control systems. A feedforward neural network (multilayer perceptron, MLP) was used as the data-driven component of the hybrid soft sensor. The network was trained on historical SCADA data using supervised learning. The model input includes physically justified features such as ΔT, as well as operational parameters of the drying system. The output of the model is the estimated moisture content $\hat{W}(t)$. This architecture ensures a balance between physical interpretability and predictive accuracy.

The proposed Big Data analysis framework forms the basis for subsequent development of hybrid soft sensors for real-time estimation of product moisture content. The hybrid soft sensor proposed in this study combines a physically based model with a data-driven correction term and can be represented as:

$$\hat{W}(t) = f_{phys}(x(t)) + f_{NN}(x(t)) \quad (1)$$

where $f_{phys}(x)$ is the analytical model based on heat and mass transfer equations, and $f_{NN}(x)$ is a neural network that compensates for unmodelled nonlinearities and disturbances; $z(t) = [T_{in}(t), T_{out}(t), \Delta T(t), G_{aIr}(t), \omega(t)]$ is the vector of instantaneous process features at time t, where T_{in} is the drying agent inlet temperature, $T_{out}(t)$ is the outlet temperature, $\Delta T(t) = T_{in} - T_{out}$ is the temperature differential, $G_{aIr}(t)$ is the airflow rate, and $\omega(t)$ is the drum rotation speed. The sliding-window feature vector is then defined as $x(t) = [z(t), z(t-1), ..., z(t-L+1)]$, where L is the sliding window length (L = 5 steps). Here, $\hat{W}(t)$ denotes the estimated product moisture content obtained from the hybrid soft sensor, while W(t) represents the actual moisture content determined from measurements. The soft sensor provides real-time estimation of W(t) based on indirectly measured process parameters.

The physical component $f_{phys}(x(t))$ encodes domain knowledge at the level of feature selection and model structure, consistent with the physics-guided (rather than physics-constrained) machine learning paradigm. Unlike physics-informed neural networks (PINNs) that enforce physical laws directly in the loss function, the present approach uses physical knowledge to motivate the choice of features (ΔT, integral thermal load, time-lagged variables) and to provide a physically interpretable baseline estimate, while the neural network correction term compensates for residual nonlinearities. Formal enforcement of physical constraints (e.g., monotonicity guarantees or non-negativity barriers on the moisture estimate) is not implemented in the current version and is left as a direction for future work.

3.1. Dataset description

The experimental data were obtained from the SCADA system of the drying department of sugar plant S (Ukraine) over two production seasons. The plant operates on a seasonal basis; the duration of the production campaign is approximately 90–110 days per year (September–December).

Data were recorded at a frequency of one record per minute, yielding a continuous raw SCADA archive of approximately 285,000–290,000 records over the two production seasons, including idle and maintenance periods between drying campaigns. From this raw archive, complete drying cycles were extracted, each spanning the intensive, stabilisation, final, and cooling stages (Figures 1–2); this yielded the modelling dataset described below (Table 1). Laboratory moisture measurements were performed using the standard oven-drying method (ICUMSA GS 2/1/3/9-15, "The Determination of Sugar Moisture by Loss on Drying") at discrete sampling points within each drying cycle, with a nominal delay of 30–60 minutes between sampling and the availability of the laboratory result. Each measurement was back-aligned to the SCADA record corresponding to the midpoint of the sampled portion. Between consecutive laboratory measurements, the target moisture value was held constant at the most recent available laboratory reading (zero-order hold / last-observation-carried-forward), rather than interpolated. This produces a piecewise-constant target series that is discontinuous at each new laboratory reading, in contrast to the continuously evolving input features; this should be taken into account when interpreting the reported accuracy metrics (Section 5), as it represents a more conservative (rather than artificially smoothed) target construction. Missing values (approximately 2.3% of records, primarily during equipment startup and shutdown) were handled by linear interpolation within gaps not exceeding 5 minutes; cycles with larger gaps were excluded. Outlier detection was performed using a per-variable 3σ criterion with a 3-record exclusion window. Each record contains the following parameters: inlet and outlet temperature of the drying agent (Tin and Tout) of the rotary drum dryer, temperature differential $\Delta T = [T_{in} - T_{out}]$, air flow rate, drum rotation speed, material feed rate, and product moisture content values determined by laboratory analysis. The dataset was divided at the cycle level into training (70%), validation (15%), and testing (15%) subsets to prevent data leakage.

Laboratory moisture measurements were used as the target variable during model training. The composition of the raw archive, the modelling dataset, and the training/validation/test subsets is summarised in Table 1; the overall data volume allows this dataset to be classified as industrial Big Data and enables the application of corresponding intelligent data analysis methods.

Table 1. Dataset composition overview

Data source / subset	Description	Size
Raw SCADA archive	Continuous process log, two production seasons, incl. idle/maintenance periods	≈285,000–290,000 records
Modelling dataset	Complete drying cycles extracted from the archive (intensive–stabilisation–final–cooling)	64 cycles / ≈18,240 records
Training subset (70%)	Cycle-level split	≈45 cycles / ≈12,800 records
Validation subset (15%)	Cycle-level split	≈9 cycles / ≈2,700 records
Test subset (15%)	Cycle-level split, held out from training	10 cycles / 2,736 records

3.2. Neural network architecture

The network comprises two hidden layers with 64 and 32 neurons respectively, both with ReLU activation functions. The input dimension is $L \times n = 25$ ($L = 5$ time steps, $n = 5$ features). Dropout regularisation (rate 0.2) was applied after each hidden layer. L2 weight decay with coefficient $\lambda = 10^{-4}$ was applied. Training used the Adam optimiser (learning rate $\eta = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with batch size 64. Maximum epochs: 200; early stopping: patience = 20 epochs on validation MSE with restoration of best-epoch weights. Input features were normalised to [0, 1] using Min-Max scaling fitted on training data; scaling parameters were stored for inference-time denormalisation. Hyperparameter selection (hidden layer sizes, dropout rate, L) was performed via grid search with 5-fold cycle-level cross-validation on the training set. Implementation: Python 3.11, TensorFlow 2.15, scikit-learn 1.4. This architecture provides a balance between computational efficiency and the ability to model the nonlinear relationships characteristic of industrial drying processes.

4. Temperature regime analysis and data preparation for iIntelligent modeling of the drying process

The sugar drying process is characterized by complex heat and mass transfer dynamics and requires the maintenance of optimal temperature regimes to ensure product quality. The primary factor determining the intensity of the drying process is the temperature regime of the drying agent.

Under industrial conditions, the drying process consists of several stages (intensive drying, stabilization, final stage, and cooling), each characterized by specific temperature regimes and moisture dynamics. The intensity of the drying process is determined by the temperature differential of the drying agent at the inlet and outlet:

$$\Delta T = T_{in} - T_{out} \tag{2}$$

where T_{in} is the temperature of the drying agent at the dryer inlet; T_{out} is the temperature of the drying agent at the outlet.

In the subsequent stage, the evaporation intensity gradually decreases as the moisture content of the product is reduced. The temperature regime stabilizes and the temperature differential diminishes.

In the final stage of drying, the product temperature approaches that of the drying agent. During this period it is important to avoid overheating, as this may lead to deterioration of product quality — in particular, caramelization and caking.

Following drying, a cooling stage is carried out during which the product temperature is reduced to values close to ambient temperature, ensuring storage stability.

The change in product moisture content during the drying process can be described by the differential equation:

$$\frac{dW}{dt} = -k \cdot (W - W_{eq}) \tag{3}$$

where W(t) is the current product moisture content; W_{eq} is the equilibrium moisture content; *k* is the drying coefficient (Zogzas et al., 1996).
The solution to this equation takes the form of an exponential dependence:

$$W(t) = W_{eq} + (W_0 - W_{eq}) \cdot e^{-kt} \tag{4}$$

where W_0 is the initial product moisture content.

To incorporate the effect of temperature conditions into the drying rate, an empirical relationship is introduced. Taking into account the technological process parameters, the drying rate coefficient k is expressed as a function of the temperature differential:

$$k = k_1 \cdot \Delta T \tag{5}$$

where k_1 is an empirical proportionality coefficient linking the drying rate to the temperature differential ΔT. This relationship reflects the proportional influence of the thermal gradient on the intensity of moisture removal and is consistent with classical heat and mass transfer principles.

Substituting (4) into (3), the moisture content is expressed as:

$$W(t) = W_{eq} + (W_0 - W_{eq}) \cdot exp(-k_1 \cdot \Delta T \cdot t) \quad (6)$$

where W(t) is the current moisture content, W_0 is the initial moisture content, W_{eq} is the equilibrium moisture content, k is the effective drying rate coefficient, k_0 is the reference drying rate coefficient, ΔT is the current temperature differential, and ΔT_{nom} is its nominal value under reference operating conditions. Thus, the drying rate coefficient k is directly proportional to the temperature differential ΔT, confirming that this parameter serves as the primary physically justified feature for soft sensor development and real-time moisture estimation. The proportionality $k \propto \Delta T$ is an empirical simplification consistent with classical convective drying theory, where the drying rate is proportional to the driving potential for heat transfer. The model parameters are estimated from training data as follows: W_0 and W_eq are identified as the mean initial and asymptotic cycle-end moisture content across training cycles (W_0 = 6.5 ±0.3%, W_{eq} = 3.0 ±0.1%); k_1 is calibrated by nonlinear least-squares minimisation of the residual between Eq. (3) and training-set moisture trajectories. Factors not captured by this simplified model — including ambient humidity, material load variability, and residence time distribution — are treated as unmodelled disturbances and compensated by the neural network correction term f_{RN}x(t)).

For subsequent use in intelligent analysis tasks, archival data obtained from SCADA/PLC systems are preprocessed. In the first step, data cleaning is performed, including the removal of missing values and anomalous readings. This is followed by synchronization of parameters by timestamp and alignment to a common time scale. Normalization of the data is then carried out to ensure the stability of machine learning algorithms. Particular attention is paid to the construction of derived features, including temperature differentials, integral characteristics, and time lags.

Table 2. Typical temperature regimes and process characteristics of the sugar drying process

Process stage	Air temperature (°C)	Product temperature (°C)	Process characteristics
Intensive drying	Elevated (within permissible limits)	60–70	Intensive moisture removal
Stabilization	Moderate	50–60	Reduction in evaporation rate
Final stage	Reduced	40–50	Prevention of overheating and preservation of product quality
Cooling	25–30	25–30	Stabilization and preparation for storage

The resulting dataset includes the following parameters: inlet and outlet air temperature, air flow rate, drum rotation speed, product feed rate, and moisture content determined by laboratory analysis.

Table 2 presents generalized temperature regimes of the sugar drying process and the corresponding characteristics of each stage, within technologically permissible values.

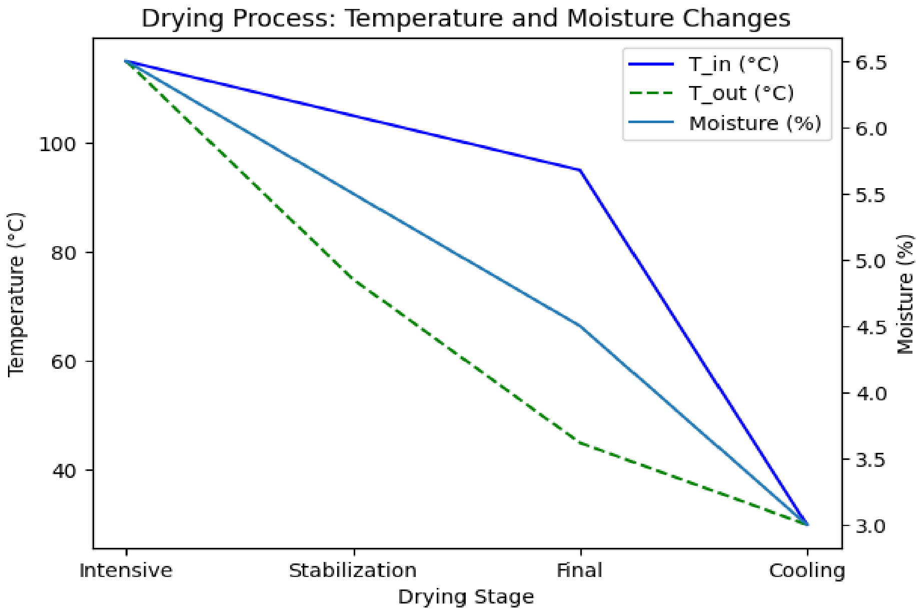

Figure 1. Change in drying agent temperature (inlet and outlet) and product moisture content at different stages of the drying process. (Based on real SCADA data from a representative test-set cycle; shown for illustration of typical process dynamics only.)

The figure shows the changes in drying agent temperature at the inlet Tin, at the outlet Tout and in product moisture content at the main stages of the drying process. In the initial stage (Intensive), the maximum inlet temperature and a significant differential between Tin and Tout are observed, corresponding to intensive heat and mass transfer accompanied by active reduction of product moisture content. In the Stabilization stage, the inlet temperature gradually decreases and the outlet temperature also falls, indicating a reduction in drying intensity and a transition to a more stabilized regime. In the Final stage, a further convergence of Tin and Tout is observed, corresponding to a slowdown in heat and mass transfer and a transition to the diffusion-limited drying stage. In the concluding Cooling stage, the inlet and outlet temperatures become nearly equal, and moisture content reaches its minimum value, indicating completion of the drying process.

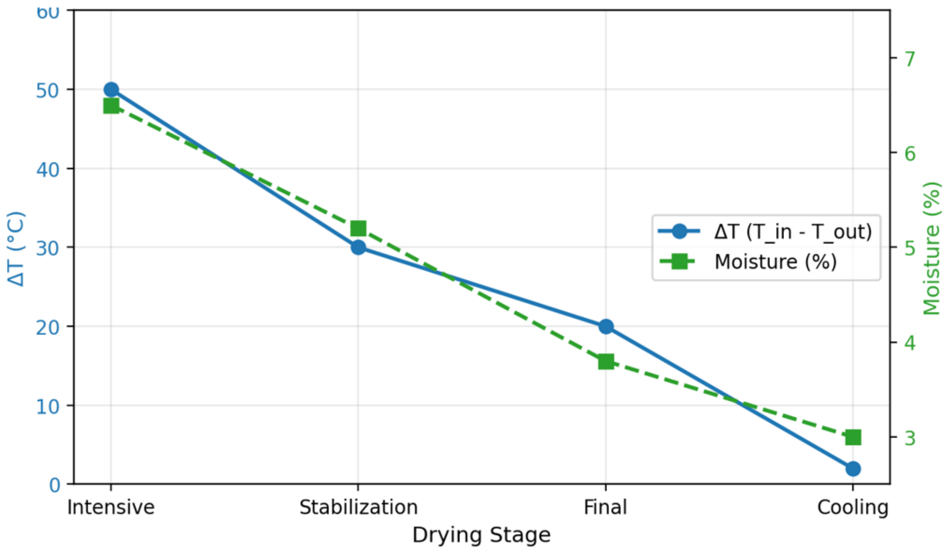

Figure 2. Dependence of the temperature differential ΔT and product moisture content at the stages of the drying process. (Based on real SCADA data from a representative test-set cycle; shown for illustration of the ΔT–moisture relationship)

Figure 2 presents the change in the temperature differential $\Delta T = T_{in} - T_{out}$ as well as the dynamics of product moisture content at the main stages of the drying process. During the intensive drying stage, ΔT reaches its maximum value (~50°C), corresponding to the most intensive moisture evaporation. As the process progresses through subsequent stages, ΔT gradually decreases to ~2°C at the cooling stage, while product moisture content decreases from 6.5% to 3.0%. The positive correlation between ΔT and moisture content confirms the physical validity of using the temperature differential as an informative feature for moisture estimation in the soft sensor.

5. Results

Table 3. Comparison of modeling approaches for product moisture content estimation

Model	MAE (%)	RMSE (%)	R^2	Training data required	Physical component
PID (indirect, temperature-based)	2.50	3.10	0.71	None	No
Linear regression (ΔT feature)	0.95	1.32	0.88	Moderate	Partial
Artificial Neural Network (ANN)	0.41	0.58	0.94	Large	No
Hybrid soft sensor (proposed)	0.062	0.089	0.97	Moderate	Yes

The test set comprised 2,736 records from 10 held-out cycles (15% split by cycle). To verify result stability, 5-fold cycle-level cross-validation was performed: MAE = 0.062 ±0.008%, RMSE = 0.089 ±0.011%, R^2 = 0.97 ±0.01. It should be noted that gravimetric loss-on-drying methods such as ICUMSA GS 2/1/3/9-15 carry an inherent measurement uncertainty arising from sample handling, drying time control, and balance resolution; the reported MAE should therefore be interpreted with this reference-method uncertainty in mind rather than as an absolute ground truth. The results presented in Table 3 demonstrate a consistent improvement in estimation accuracy as the model complexity and physical grounding increase.

Note: all four approaches in Table 3 were evaluated on the identical test set comprising 10 held-out cycles (2,736 records, 15% of the modelling dataset — see Table 1 for the full dataset composition — never seen during training). The PID entry represents an indirect moisture estimate derived from the temperature control signal via the inverse of Eq. (3) at nominal operating conditions; it serves as a process-engineering baseline rather than a learned estimator, and its inclusion clarifies the accuracy gap between temperature-based control and data-driven approaches. The baseline PID controller, which relies on indirect temperature-based regulation, yields the lowest accuracy (R^2 = 0.71), reflecting the well-known limitations of open-loop moisture control in industrial dryers. The absence of direct product quality feedback introduces systematic estimation errors under varying raw material properties and load conditions.

The linear regression model incorporating the physically justified feature $\Delta T = T_{in} - T_{out}$ achieves a notably higher coefficient of determination (R^2 = 0.88), confirming that the temperature differential carries significant predictive information about the drying state. However, the linear formulation is insufficient to capture the nonlinear dynamics of the evaporation process, particularly during the transition between drying stages.

The standalone ANN model achieves R^2 = 0.94 by exploiting nonlinear relationships in the training data. Nevertheless, without explicit physical constraints, the model may produce physically inconsistent predictions outside the training distribution, limiting its reliability under process disturbances.

The proposed hybrid soft sensor combines the physical drying model with neural network correction and achieves the best performance across all metrics: MAE = 0.062%, RMSE = 0.089%, R^2 = 0.97. The integration of physically justified features ensures model interpretability and stability, while the neural network component compensates for residual nonlinearities and measurement uncertainties not captured by the analytical model (Perré et al., 2025).

Figure 3 presents a comparison of the actual product moisture content W(t) with the values predicted by the hybrid soft sensor $\hat{W}(t)$ and the standalone ANN model over a complete drying cycle. The hybrid model demonstrates high accuracy in reproducing moisture dynamics across all process stages — from intensive drying to cooling. The deviation of predicted values from actual values remains within MAE = 0.062%, confirming the model adequacy and its ability to reproduce the nonlinear dynamics of the process. The ANN model also provides acceptable accuracy (MAE = 0.41%); however, it exhibits larger deviations at transition zones between drying stages, which is attributed to the absence of physical constraints in a purely data-driven approach.

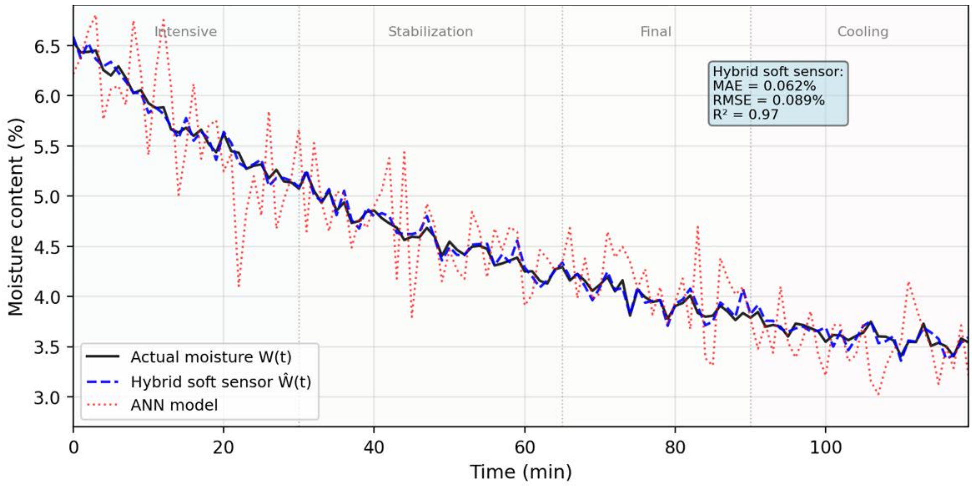

Figure 3. Comparison of actual and predicted moisture content during the sugar drying process. (Based on real test-set data; model predictions obtained on held-out cycles not used during training)

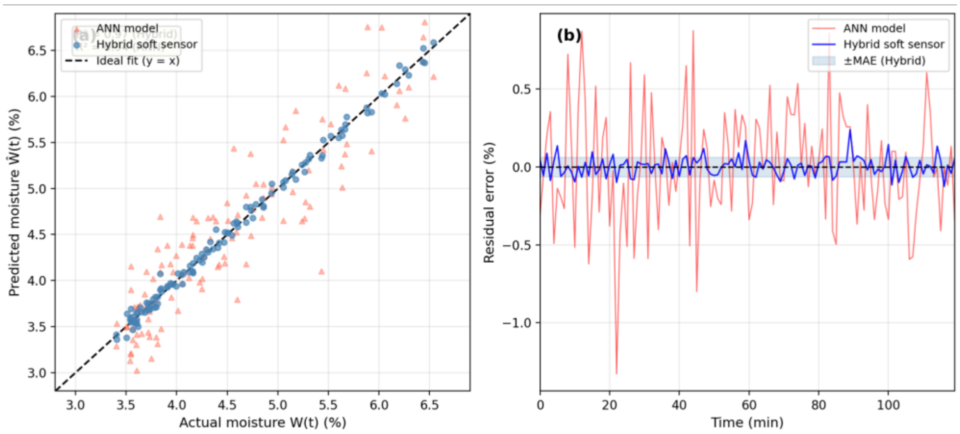

Figure 4. Predicted vs actual moisture content (scatter plot) and prediction residuals over time. (Aggregated over all 2,736 test-set records; constitutes quantitative evidence of model performance on unseen data). **Figure 4a.** Scatter plot of predicted moisture content $\hat{W}(t)$ versus actual moisture content W(t) for the hybrid soft sensor and the standalone ANN model. The hybrid model data points (blue circles) are tightly clustered along the ideal fit line y = x across the full moisture range of 3.0–6.5%, confirming the high prediction accuracy ($R^2 = 0.97$). The ANN model data points (pink triangles) show noticeably greater scatter, particularly at moisture values below 4.5%, corresponding to the final drying stage and cooling. **Figure 4b.** Prediction residuals over time for the hybrid soft sensor (blue line) and the ANN model (red line) during a complete drying cycle (~110 min). The hybrid model residuals remain uniformly distributed within the ±MAE band (light blue shading, ±0.062%), with no visible systematic bias throughout the entire cycle. The ANN model residuals exhibit significantly higher amplitude (up to ±1.0%) and show irregular oscillations, indicating the inability of the purely data-driven approach to fully capture the physical dynamics of the drying process.

5.1. Ablation study

To assess which components of the hybrid architecture are responsible for the observed accuracy improvement, an ablation study was conducted on the same test set (Table 4). Four configurations were evaluated: (A) physical model alone — f_{phys}(x(t) without neural network correction (R^2 = 0.88, MAE = 0.31%); (B) MLP alone — trained to predict full moisture content without the physical baseline (R^2 = 0.94, MAE = 0.41%); (C) hybrid model without time-lagged features — f_{phys}+f_{NN}using only current-time variables (L = 1) (R^2 = 0.93, MAE = 0.18%); (D) full hybrid (proposed) — f_{phys}+$f_{NN}$with sliding window L = 5 (R^2 = 0.97, MAE = 0.062%). These results demonstrate that each component contributes meaningfully: adding the neural network correction to the physical baseline (without time lags) increases R^2 from 0.88 to 0.93 — an absolute increase of 5 percentage points (a relative improvement of approximately 5.7%) (configuration (C) vs. (A)) — while incorporating time-lagged features further increases R^2 from 0.93 to 0.97, an absolute increase of 4 percentage points (a relative improvement of approximately 4.3%) (configuration (D) vs. (C)). It should be noted that MAE and R^2 do not always rank configurations identically: configuration (A) achieves a lower MAE than (B) despite a lower R^2, and the same pattern recurs between (B) and (C), where MAE decreases substantially (0.41% to 0.18%) while R^2 decreases slightly (0.94 to 0.93). This likely reflects the different sensitivities of the two metrics: R^2 depends on the variance explained across the full test-set range, whereas MAE reflects the average error magnitude, so a model can reduce typical errors while explaining a marginally smaller share of the overall variance.

Table 4. Ablation study: contribution of individual model components (test set, 10 held-out cycles)

Configuration	MAE (%)	RMSE (%)	R^2
(A) Physical model alone	0.31	0.43	0.88
(B) MLP alone	0.41	0.58	0.94
(C) Hybrid, no time lags (L = 1)	0.18	0.25	0.93
(D) Full hybrid — proposed	**0.062**	**0.089**	**0.97**

5.2. Limitations

Despite the high accuracy of the developed model, several limitations should be acknowledged when considering its practical application.

First, the model was trained on data from a single type of equipment — the rotary drum dryer of sugar plant S. Transfer of the model to other types of drying equipment or facilities with different process parameters requires retraining or adaptation of the model.

Second, the equilibrium moisture content W_{eq} in the physics-based component of the model is treated as a constant. Under real operating conditions, this parameter may vary depending on raw material properties, ambient humidity, and seasonal factors, which may introduce additional error into the moisture estimation.

Third, model validation was performed on archival data in offline mode. Real-time testing under actual production conditions was not conducted, which represents a necessary step prior to industrial deployment.

Fourth, laboratory moisture measurements used as the target variable were performed with a certain time delay relative to the process parameters. This may introduce synchronisation error into the training dataset.

Fifth, the current implementation does not include a closed-loop retraining mechanism: the model is trained offline on historical data and deployed as a static estimator. Development of a continuous adaptation loop, in which the soft sensor is periodically retrained as new labelled data become available, is left as a direction for future work.

Addressing these limitations is the subject of further research, including real-time model testing and dataset expansion through data from multiple production campaigns.

5.3. Practical implementation

The proposed hybrid soft sensor was designed with the possibility of integration into existing industrial automation systems without significant modification of the plant's hardware infrastructure.

The implementation architecture is as follows. The SCADA system collects process parameters in real time — inlet and outlet drying agent temperatures (Tin, Tout), air flow rate, drum rotation speed, and material feed intensity. These data are transmitted to the soft sensor computational module, implemented as a standalone software component. The module computes an estimate of the current moisture content $\hat{W}(t)$ based on the hybrid model and returns the result to the SCADA system as a virtual tag.

The resulting value $\hat{W}(t)$ can be used by the operator as an indicator of the current process state, as well as an input signal for the automatic control system — in particular, a PID controller for drying agent temperature or flow rate. This enables a transition from fixed process regimes to adaptive real-time control of the drying process.

The advantage of this approach is that implementation does not require the installation of additional inline moisture analysers, which significantly reduces capital costs. The soft sensor effectively replaces an expensive online moisture analyser by utilising only the existing SCADA signals.

Practical implementation requires three steps: configuring data transfer from the SCADA system to the computational module (via OPC UA or Modbus protocol), deploying the trained model on an industrial PC or cloud service, and configuring the feedback loop as a virtual moisture tag within the SCADA system.

6. Conclusions

A process-oriented approach to modeling and control of the sugar drying process based on historical SCADA data has been developed. Physics-guided feature construction from archival SCADA data, including the temperature differential ΔT, is described in Section 3 and validated in Section 4 (Eqs. 2–6).

K-means clustering identifying four operating regimes is presented in Section 3.

A hybrid soft sensor combining a physical model with neural network correction has been proposed for real-time estimation of product moisture content. The developed model demonstrated high prediction accuracy on the held-out test set (MAE = 0.062%, RMSE = 0.089%, R^2 = 0.97), outperforming PID control, linear regression, and a standalone ANN on the investigated dataset, as reported in Section 5 (Table 3).

The ablation study quantifying the contribution of each model component is presented in Section 5.1 (Table 4), showing that both the physical baseline and the time-lagged features contribute meaningfully to the overall accuracy.

The proposed approach can be implemented in existing industrial systems and serves as a foundation for the further development of adaptive and intelligent drying control systems.

The prospects for further research lie in extending the approach to different types of drying equipment, integrating the model with advanced control methods including adaptive and predictive control, and developing a closed-loop retraining mechanism for continuous model adaptation as new labelled data become available.

List of Abbreviations

ANN - Artificial Neural Network
MAE - Mean Absolute Error
MLP - Multilayer Perceptron
MSE - Mean Squared Error
PID - Proportional-Integral-Derivative (controller)
PLC - Programmable Logic Controller
RMSE - Root Mean Square Error
SCADA - Supervisory Control and Data Acquisition

References

Fernández-Caramés, T. M., Fraga-Lamas, P., Suárez-Albela, M., Castedo, L. (2021). Sensors data analysis in SCADA systems to foresee failures. *Sensors*, 21(8), 2762. https://doi.org/10.3390/s21082762

Gryhorchuk, H. V., Gryhorchuk, L. I., Khrabatyn, R. I. (2025). Mathematical modelling of massecuite drying process in a drum dryer. *Technical Sciences and Technologies*, 2(52), 63–70. https://doi.org/10.25140/2411-5363-2025-2(52)-63-70

Kadlec, P., Gabrys, B., Strandt, S. (2009). Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, 33(4), 795–814. https://doi.org/10.1016/j.compchemeng.2008.12.012

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168

Khan, M. I. H., Batuwatta-Gamage, C. P., Karim, M. A., Gu, Y. (2022a). Fundamental understanding of heat and mass transfer processes for physics-informed machine learning-based drying modelling. *Energies*, 15(24), 9347. https://doi.org/10.3390/en15249347

Khan, M. I. H., Sablani, S. S., Joardder, M. U. H., Karim, M. A. (2022b). Application of machine learning-based approach in food drying: Opportunities and challenges. *Drying Technology*, 40(6), 1051–1067. https://doi.org/10.1080/07373937.2020.1853152

Kudra, T., Mujumdar, A. S. (2009). *Advanced drying technologies* (2nd ed.). CRC Press.

Kusiak, A. (2017). Smart manufacturing must embrace big data. *Nature*, 544(7648), 23–25. https://doi.org/10.1038/544023a

Miraei Ashtiani, S. H., Martynenko, A. (2024). Toward intelligent food drying: Integrating artificial intelligence into drying systems. *Drying Technology*, 42(8), 1240–1269. https://doi.org/10.1080/07373937.2024.2356177

Mujumdar, A. S. (Ed.). (2015). *Handbook of industrial drying* (4th ed.). CRC Press.

Perré, P., Rémond, R., Turner, I. W., Brochard, L. (2025). Toward mechanistic models 'augmented' by machine learning. *Drying Technology*, 43(1–2), 147–161. https://doi.org/10.1080/07373937.2024.2411582

Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control*, 36(2), 220–234. https://doi.org/10.1016/j.arcontrol.2012.09.004

von Stosch, M., Oliveira, R., Peres, J., Feyo de Azevedo, S. (2014). Hybrid semi-parametric modeling in process systems engineering. *Computers & Chemical Engineering*, 60, 86–101. https://doi.org/10.1016/j.compchemeng.2013.08.008

Wang, L., Törngren, M., Onori, M. (2015). Current status and advancement of cyber-physical systems in manufacturing. *Journal of Manufacturing Systems*, 37(2), 517–527. https://doi.org/10.1016/j.jmsy.2015.04.008

Zogzas, N. P., Maroulis, Z. B., Marinos-Kouris, D. (1996). Moisture diffusivity data compilation in foodstuffs. Drying Technology, 14(10), 2225–2253. https://doi.org/10.1080/07373939608917205

Received May 6, 2026, revised August 18, 2026, accepted September 20, 2026