ICA2025OSLO
14th IWA Specialised Conference on
Instrumentation, Control and Automation

# Impact of Data Dimensionality Reduction with Autoencoders on Predictive Models for Nitrogen Conversion Rates in Wastewater Treatment Denitrification Processes

A.E. Voipan *, D. Voipan **, M. Barbu***

*Department of Automation, Faculty of Automation, Computers, Electrical Engineering and Electronics, 'Dunarea de Jos' University of Galati, Romania, andreea.voipan@ugal.ro

**Department of Computer Science, Faculty of Automation, Computers, Electrical Engineering and Electronics, 'Dunarea de Jos' University of Galati, Romania, daniel.voipan@ugal.ro

***Department of Automation, Faculty of Automation, Computers, Electrical Engineering and Electronics, 'Dunarea de Jos' University of Galati, Romania, marian.barbu@ugal.ro

**Abstract:** This study investigates the impact of dimensionality reduction on deep learning models for predicting nitrogen conversion rates (NCRs) in wastewater treatment plants (WWTPs), focusing on denitrification dynamics. Using simulation data from the Benchmark Simulation Model No. 2 (BSM2), three input representations were explored: (1) the full set of seven operational variables, (2) a reduced subset of four features selected via Random Forest regression, and (3) a four-dimensional latent representation extracted through a deep Autoencoder network. Each configuration was used to train a Long Short-Term Memory (LSTM) network for forecasting effluent concentrations of nitrate ($SNO_{out}$) and ammonium ($SNH_{out}$). The model trained on Autoencoder-compressed features (AE_4_latent) achieved the best predictive performance (MSE: 0.0006, $R^2$: 0.9609), outperforming both the full-input model ($R^2$: 0.9475) and the Random Forest-based model ($R^2$: 0.9320), while requiring fewer input dimensions and maintaining similar GPU memory usage. All trained models were successfully exported to ONNX format and deployed in MATLAB/Simulink for real-time validation within the BSM2 environment. These results confirm that nonlinear dimensionality reduction using Autoencoders can enhance both prediction accuracy and computational efficiency, making this approach highly suitable for real-time control applications in WWTPs.

**Keywords**: Autoencoder; dimensionality reduction; WWTP; nitrogen conversion rates (NCR); LSTM; RF.

## 1. INTRODUCTION

Predictive models are increasingly being recognized as powerful tools for improving the performance and operational efficiency of wastewater treatment plants (WWTPs), particularly in the context of biological nitrogen (N) removal processes such as nitrification and denitrification. Accurate prediction of nitrogen conversion rates (NCRs), which represent the transformation of nitrogenous compounds such as ammonium ($NH_4^+$), nitrate ($NO_3^-$), nitrite ($NO_2^-$), and nitrous oxide ($N_2O$), is essential for effective process control, energy optimization, and emission reduction strategies in WWTPs [Foley et al., 2010].

However, WWTPs generate complex, high-frequency, multivariate time-series data composed of influent characteristics (chemical oxygen demand (COD), $NH_4^+$, alkalinity), operational parameters, such as temperature, dissolved oxygen (DO), total suspended solids (TSS), and internal process dynamics. The high dimensionality of such datasets presents significant challenges to predictive modelling: not only does it increase computational costs and training time, but it also reduces model interpretability and increases the risk of overfitting.

To address this, dimensionality reduction techniques have emerged as promising tools to extract compact, informative representations of the input space. In particular, Autoencoders (AEs), deep neural networks (DNNs) designed to learn compressed representations, have demonstrated their ability to reduce feature space while retaining key information necessary for accurate forecasting [Wang et al., 2021]. Additionally, classical machine learning (ML) methods such as Random Forests (RF) offer the benefit of feature importance ranking, enabling interpretable selection of the most relevant input variables.

This study evaluates and compares three Long Short-Term Memory (LSTM)-based architectures for forecasting NCRs using simulation data from the Benchmark Simulation Model No. 2

(BSM2), a widely accepted framework for WWTP control strategies [van Loosdrecht and Brdjanovic, 2014]. The three approaches are:

- An LSTM trained on all 7 original variables from the BSM2 dataset (baseline);
- An LSTM trained on a reduced subset of 4 variables selected via RF feature importance;
- An LSTM trained on 4 latent features generated by a deep AEs network.

All three networks are evaluated in terms of prediction accuracy (MSE, MAE, $R^2$) and computational efficiency (GPU memory usage). Furthermore, each trained LSTM model was first saved in Keras format, then converted to ONNX, and finally integrated into MATLAB/Simulink as a. mat-based NN block using the Deep Learning Toolbox, enabling real-time simulation and comparative validation within the BSM2 environment.

The central objective of this paper is to demonstrate that dimensionality reduction, whether linear (RF) or nonlinear (AE), can enhance both the accuracy and practical applicability of LSTM networks in WWTP predictive tasks, thus providing a robust, efficient foundation for future real-time control implementations.

## 2. DATA SOURCE AND PREPROCESSING

The dataset used in this study was generated using the BSM2, a well-established simulation framework for evaluating operational and control strategies in wastewater treatment plants (WWTPs) [Jeppsson, 2007]. BSM2 provides realistic influent dynamics and process responses, covering a simulation period of 609 days with data sampled at 15-minute intervals, resulting in approximately 58,465 valid time steps after filtering. The selected variables, illustrated in Figure 2.1, include seven influent and operational parameters and three target effluent concentrations, all of which are critical to characterizing NCRs during biological treatment.

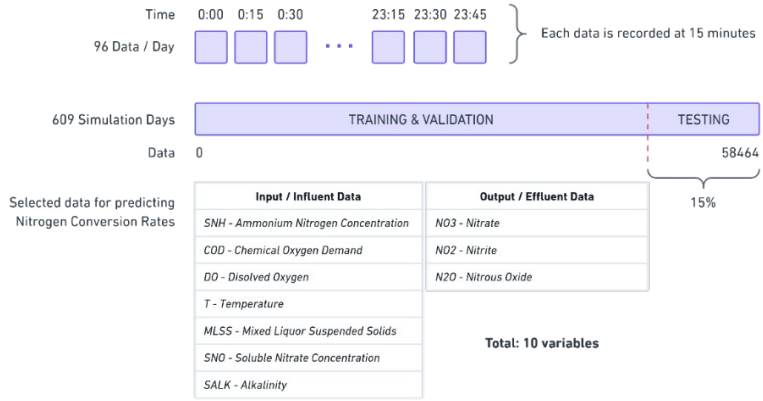


**Figure 2.1** Dataset structure for training, validation, and testing in predicting nitrogen conversion rates in wastewater treatment.

To ensure data consistency, all missing or undefined entries were removed. The remaining data was scaled using Min-Max normalization to the [0, 1] interval, and the scaling parameters were saved to support consistent preprocessing during Simulink deployment. To assess the impact of dimensionality reduction, three distinct preprocessing pathways were designed: one using the full set of seven input variables, another based on four features selected through RF importance scores, and a third using four latent features extracted via an AE network. These parallel branches are illustrated in Figure 2.2 (a).

Each version of the input was then transformed into supervised sequences using a sliding window: 40-time steps (10 hours) were used as input to predict the next 16 steps (4 hours) [Țîru et al., 2023], a structure that enables LSTM networks to learn temporal dependencies, as depicted in Figure 2.2 (b).

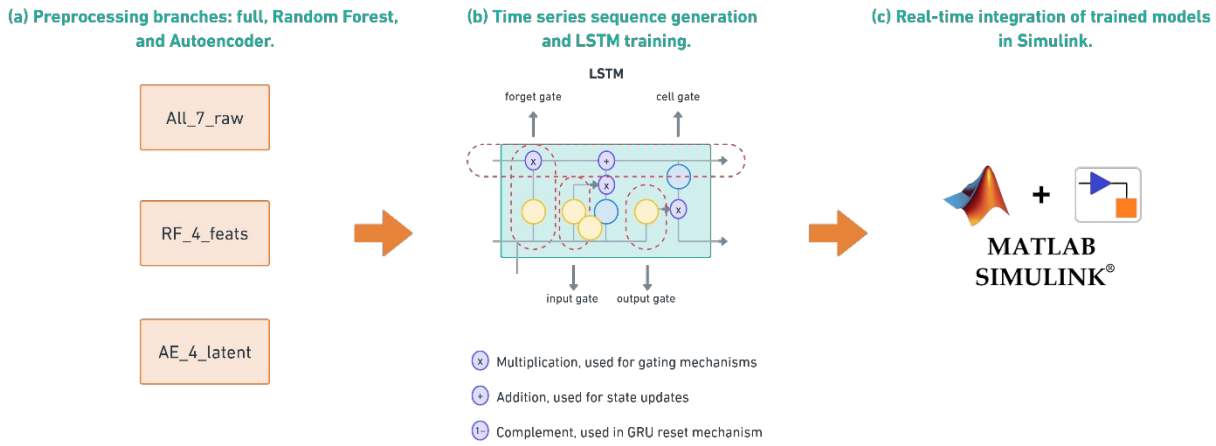


**Figure 2.2** Workflow for LSTM modelling and Simulink integration.

Finally, the trained LSTM models were exported as ONNX files and integrated into MATLAB/Simulink, where they were connected to real-time simulation outputs from BSM2. The full deployment and runtime prediction logic are summarized in Figure 2.2 (c).

## 3. METHODOLOGY

### 3.1 Autoencoders for dimensionality reduction

To reduce the input dimensionality while preserving critical process information, a deep AE NN was implemented. This architecture consists of an encoder, which compresses the original 7 input variables into a 4-dimensional latent space, and a decoder, which reconstructs the input from this compact representation [Goodfellow et al., 2016]. The encoder and decoder were implemented as symmetrical fully connected layers. The model was trained using the Adam optimizer (learning rate (LR) = 0.001) and MSE loss, with early stopping and dynamic LR reduction to enhance generalization [Voipan et al., 2025]. After training over 50 epochs, the AE achieved a reconstruction MSE of 0.0001, a MAE of 0.0036, and an $R^2$ of 0.5840, confirming that the latent space retained the most relevant information with minimal distortion.

The resulting compressed features were used to train the AE_4_latent LSTM model, Figure 2.2 (a), which showed the highest overall predictive performance. This confirms that AE not only reduce data dimensionality but also enable efficient and accurate temporal modeling, especially in WWTP applications, as also explored in [Lee et al., 2021].

### 3.2 Random Forest feature selection

As an alternative to deep, unsupervised dimensionality reduction, this study also explored a feature selection approach using RF regression models. RF is an ensemble learning method that constructs multiple decision trees and combines their predictions, offering both high accuracy and interpretability via feature importance scores [Breiman, 2001]. In this context, two RF regressors were trained separately to predict $SNO_{out}$ and $SNH_{out}$ based on the full set of seven normalized input variables. Feature importance was computed based on how much each variable reduced prediction error across the decision trees.

The analysis revealed that temperature, SNH, TSS, and COD were consistently the most informative features for both outputs. In contrast, $SO_{in}$, $SNO_{in}$, and SALK had minimal influence. As a result, the top four features were selected for training the reduced-input LSTM model, referred to as RF_4_feats in Figure 2.2 (b).

Despite its simplicity, the RF-based feature set yielded solid performance: $R^2$ values exceeded 0.94 for both targets, and the models achieved MSEs of 0.0536 ($SNO_{out}$) and 0.0546 ($SNH_{out}$).

While this approach preserves interpretability and reduces input dimensionality, it lacks the nonlinear compression and abstraction capabilities of autoencoders, which may better capture hidden dependencies in the data.

### 3.3 LSTM Network Architectures and Training Setup

To evaluate the effect of dimensionality reduction on temporal prediction performance, three distinct LSTM architectures were developed and trained using different versions of the input dataset: the full 7-variable input set (all_7_raw), a 4-variable subset selected using RF (RF_4_feats), and a 4-dimensional latent representation extracted via AE (AE_4_latent). Input sequences were generated with a sliding window, and data was split using an 80/20 train-test ratio. A summary of the model configurations is presented in Table 3.1.

All models were trained with early stopping and LR scheduling to improve generalization and reduce training time. As presented in the results section, the (AE_4_latent) model achieved the best performance, combining the lowest MSE, highest $R^2$, and lowest GPU memory usage. These results confirm the advantages of nonlinear dimensionality reduction for deep temporal modelling in WWT applications.

**Table 3.1** Model configurations and input strategies

| Aspect | all_7_raw | RF_4_feats | AE_4_latent |
|---|---|---|---|
| **Input dimension** | 7 raw variables | 4 RF-selected variables | 4 compressed latent features |
| **LSTM layers** | 2 layers (256, 128 units) | | |
| **Dense layers** | 32 neurons + output layer | | |
| **Activations** | tanh/sigmoid (LSTM), ReLU (Dense) | | |
| **Optimizer / loss** | Adam, lr = 0.001; MSE | | |
| **Epochs / batch size** | 50 / 64 | | |
| **Prediction horizon** | 16 steps (4 hours) | | |
| **Training input** | Full scaled input | Reduced input (RF-selected) | Latent features from AE |

### 3.4 Integration in Simulink

To validate the trained models in a dynamic, real-time environment, the integration pipeline, illustrated in Figure 3.1, consists of three ONNX-based Simulink blocks (one for each LSTM variant), with input normalization performed using .mat files containing the MinMaxScaler parameters from training. Each model receives synchronized influent data from BSM2, processes it in parallel, and returns denormalized predictions in physical units. The outputs, out.all_SNO, out.rf_SNO, and out.ae_SNO, allow direct comparison of forecasting accuracy under identical simulation conditions.

To ensure end-to-end consistency, the same normalization logic used during offline training was preserved and applied within Simulink. Each model processes the scaled inputs independently and returns real-time predictions of $SNO_{out}$ and $SNH_{out}$, which are then used for performance comparison and potential control integration.

This modular and scalable integration confirms the practical applicability of Python-trained DL models in MATLAB/Simulink environments. It also offers a reproducible platform for real-time

evaluation, sensitivity analysis, or future extensions to control-oriented architectures like model predictive control (MPC).

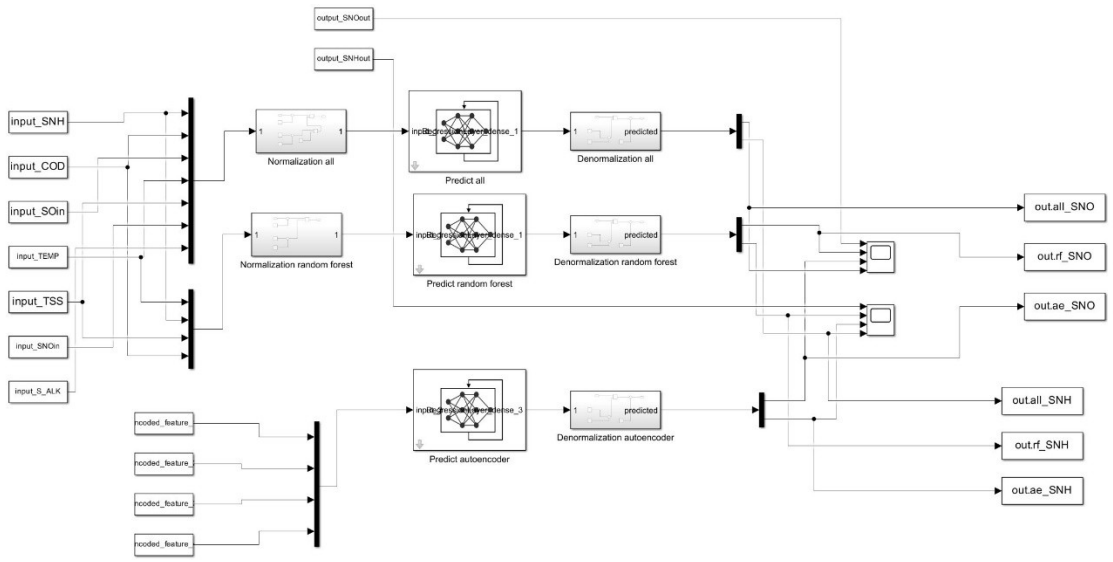


**Figure 3.1** Simulink integration of LSTM models for real-time prediction of N concentrations.

## 4. RESULTS AND DISCUSSION

### 4.1 Predictive Performance

The predictive accuracy of the LSTM models was evaluated using three standard metrics: MSE, MAE, and $R^2$ [Voipan et al., 2024]. As shown in Table 4.1, the model trained on AE-compressed features (AE_4_latent) achieved the best overall performance, with an MSE of 0.0006, MAE of 0.0161, and $R^2$ of 0.9609. The baseline model using all seven raw inputs (all_7_raw) also performed strongly, with $R^2 = 0.9475$, while the RF-based model (RF_4_feats) demonstrated slightly lower accuracy ($R^2 = 0.9320$).

**Table 4.1** Performance metrics for LSTM models trained on different input representations

| Model Name | Input Type | MSE | MAE | $R^2$ |
|---|---|---|---|---|
| **All_7_raw** | 7 raw input variables | 0.0008 | 0.0191 | 0.9475 |
| **RF_4_feats** | 4 RF-selected features | 0.0011 | 0.0218 | 0.9320 |
| **AE_4_latent** | 4 latent features from Autoencoder | 0.0006 | 0.0161 | 0.9609 |

These results confirm that nonlinear dimensionality reduction via autoencoders not only preserves essential temporal patterns but can also enhance generalization, despite using fewer input features. This supports previous findings in the literature [Lee et al., 2021], where compressed representations led to better learning outcomes in temporal models.

To better visualize the quality of predictions, Figure 4.1 illustrates a representative time series interval comparing actual and predicted values of SNH and SNO for both RF_4_feats and AE_4_latent models. The Autoencoder-based model exhibits closer alignment with true values, especially during peak variations in SNH and plateau regions in SNO. Zoomed-in views further emphasize the improved temporal resolution achieved through latent feature compression.

To assess prediction consistency across the entire dataset, Figure 4.2 presents a global time-series comparison of outputs from all three LSTM variants. The predictions from AE_4_latent

closely track those from the full input model, while those from RF_4_feats show more deviation during transient states. These results reinforce the capability of nonlinear compression via autoencoders to preserve critical temporal information.

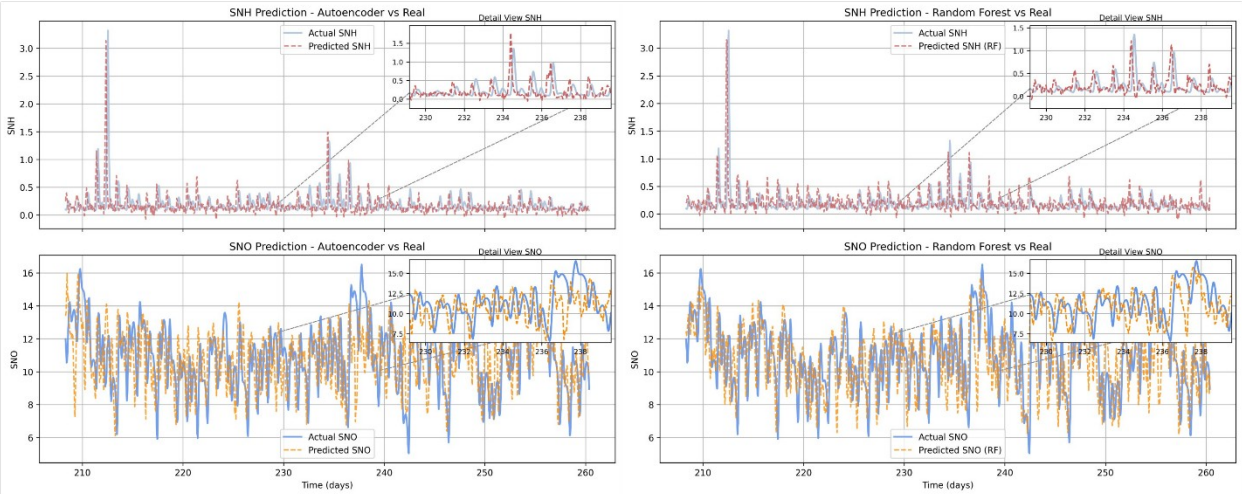


**Figure 4.1** Predicted vs. actual SNH and SNO concentrations using RF and AE-based LSTM models.

### 4.2 Computational efficiency

In addition to accuracy, computational efficiency was evaluated based on GPU memory usage and input dimensionality. All models had similar peak memory requirements (approx. 97–98 MB), but models trained on reduced inputs (RF_4_feats and AE_4_latent) benefit from faster inference and lower memory demands in real-time applications.

Furthermore, the AE model reduces not only the number of inputs but also the input correlation noise, producing smoother predictions (as visible in Figure 4.1). This makes it particularly suitable for use in resource-constrained deployments or for future integration into adaptive control architectures.

### 4.3 Feature interpretability

Feature importance analysis via RF revealed that temperature and SNH were the most predictive inputs for NCR outputs, followed by TSS and COD. As mentioned in 3.2 Random Forest feature selection, other variables such as SOin, SNOin, and SALK had negligible influence.

This insight enables sensor prioritization and process simplification in operational WWTP settings. While AE lack direct interpretability, they provide an abstracted and robust representation of input dynamics, suitable for systems where interpretability is secondary to performance.

### 4.4 Real-time integration results

All three LSTM models were successfully exported to ONNX format and deployed in Simulink for real-time execution within the BSM2 environment. As illustrated in Figure 3.1, each model was encapsulated as a NN block and received normalized inputs directly from BSM2 outputs. The system returned synchronized forecasts of SNOout and SNHout in real time (Figure 4.2), via outputs out.all_SNO, out.rf_SNO, and out.ae_SNO.

In this live setting, all models tracked the system behavior effectively. However, the AE_4_latent model exhibited smoother transitions and better responsiveness to short-term

fluctuations, likely due to noise filtering in the compressed latent space. This characteristic makes it a strong candidate for integration into MPC loops and early warning systems in WWTP operation.

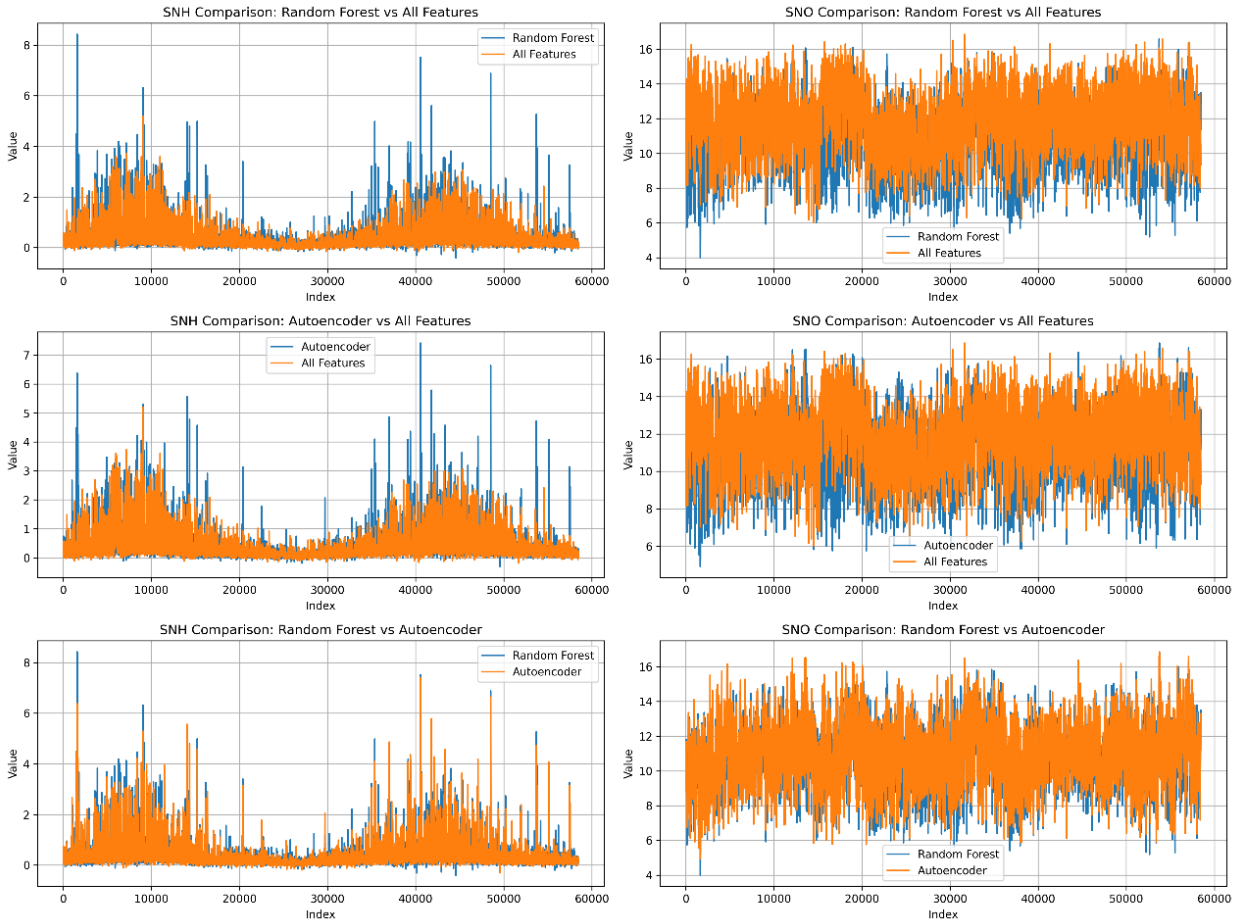


**Figure 4.2** Global comparison of predicted SNH and SNO concentrations across all LSTM models.

## CONCLUSIONS

This study evaluated the performance of LSTM neural networks trained on three different input representations, raw operational data, features selected via Random Forest, and compressed latent features generated by an Autoencoder, applied to the prediction of nitrogen compound concentrations in a simulated wastewater treatment process, BSM2.

The results highlight the practical and methodological value of dimensionality reduction in improving both model performance and deployment efficiency. Among the three configurations tested, the LSTM model trained on AE-compressed features (AE_4_latent) consistently outperformed the other variants, achieving the highest predictive accuracy ($R^2$ = 0.9609) and lowest error metrics (MSE = 0.0006). Despite operating on only four latent variables, it surpassed the model trained on the full 7-variable input space, demonstrating the effectiveness of nonlinear compression in preserving essential temporal patterns.

The RF-based feature selection (RF_4_feats) offered interpretability and reduced model size, but with a slight compromise in performance. However, it remains a viable option for applications where transparency and explainability are prioritized.

All models were successfully exported in ONNX format and integrated into Simulink for real-time inference within the BSM2 framework. The integration process confirmed that DL models developed in Python can be reliably deployed in MATLAB/Simulink environments, opening the door for broader adoption in process control and digital twin systems.

Future work will focus on several extensions. First, integrating explainable AI (XAI) techniques into the AE-LSTM pipeline could improve the interpretability of compressed features. Second, coupling the predictive model with real-time control loops (for aeration or carbon dosing) could enhance operational efficiency. Third, testing the approach on real-world WWTP datasets would allow validation under non-simulated noise and disturbances. Finally, evolving the architecture toward hybrid physics-informed models could bridge the gap between data-driven and mechanistic approaches, improving robustness in unseen scenarios.

## ACKNOWLEDGEMENT

This work was supported by a grant from the Ministry of Education and Research of Romania, project number PN-IV-P7-7.1-PED-2024-0910, under the National Plan for Research, Development and Innovation (PNCDI IV).

## REFERENCES

**[Foley et al., 2010]** Foley, J. A., et al. (2010). Global consequences of land use. Science, 290(5495), 1202-1204.

**[Goodfellow et al., 2016]** Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

**[Jeppsson, 2007]** Jeppsson, U. (2007). Benchmark simulation model no 2 for wastewater treatment plant design and operation. Journal of Environmental Engineering, 133(10), 1010-1018.

**[Lee et al., 2021]** Lee, D., Jeon, J., & Park, C. (2021). Deep Learning for Predicting Nitrogen Removal in Wastewater Treatment. Environmental Science and Technology, 55(3), 1283-1290.

**[Țîru et al., 2023]** Țîru, A.E., Vasiliev, I., Diaconu, L., Vilanova, R., Voipan, D., & Ratnaweera, H. 2023. Integration of ANN for Accurate Estimation and Control in Wastewater Treatment. *Proceedings of the IEEE 28th International Conference on Emerging Technologies and Factory Automation (ETFA)*, Sinaia, Romania, pp. 1–4.

**[van Loosdrecht and Brdjanovic, 2014]** van Loosdrecht, M., & Brdjanovic, D. (2014). Wastewater Treatment: Biological and Chemical Processes (2nd ed.). Springer.

**[Voipan et al., 2024]** Voipan, D., Vasiliev, I., Diaconu, L., Voipan, A.-E., & Barbu, M. 2024. Enhancing Effluent Quality Predictions in Wastewater Treatment with LSTM Neural Network. *Proceedings of the 28th International Conference on System Theory, Control and Computing (ICSTCC)*, Sinaia, Romania, pp. 83–88. DOI: 10.1109/ICSTCC62912.2024.10744668

**[Voipan et al., 2025]** Voipan, D., Voipan, A.E., & Barbu, M. 2025. Evaluating Machine Learning-Based Soft Sensors for Effluent Quality Prediction in Wastewater Treatment Under Variable Weather Conditions. *Sensors*, 25(6), 1692. DOI: 10.3390/s25061692

**[Wang et al., 2021]** Wang, K., Chang, P., & Meng, F. (2021). Monitoring of Wastewater Treatment Process Based on Slow Feature Analysis Variational Autoencoder. In *Proceedings of the 2021 IEEE 10th Data Driven Control and Learning Systems Conference (DDCLS)*, Suzhou, China, 2021, 495-502.