eISSN 2783-6851

VILNIUS TECH
Vilnius Gediminas Technical University

NEW TRENDS in COMPUTER SCIENCES

2025
Volume 3
Issue 1
Pages 38–48
https://doi.org/10.3846/ntcs.2025.24139

EFFECT OF NOISE REDUCTION ON PLSR MODELING IN NEAR INFRARED SPECTROSCOPY USING DENOISING AUTOENCODER

Özcan ÇATALTAŞ

Department of Electrical and Electronic Engineering, Selcuk University, Konya, Türkiye

Article History:
- received 11 June 2025
- accepted 18 June 2025

Abstract. In this study, a deep learning-based denoising autoencoder approach is proposed to increase the robustness of near-infrared spectroscopy data to random noise and improve quantitative modeling accuracy. Artificial Gaussian noise at four different levels (10, 15, 20, and 25 dB) was added to the near-infrared spectra obtained from milk samples to mimic the real measurement conditions. The noisy spectra were denoised by processing with an autoencoder architecture consisting of fully connected layers. The noise removal performance is quantitatively evaluated with both theoretical and measured signal-to-noise ratio values. The results show that the AE model significantly improves the spectral signal quality at all signal-to-noise ratio levels. In particular, at the lowest signal-to-noise ratio level (10 dB), the signal-to-noise ratio value nearly tripled to 29.6 dB with the autoencoder. At all other levels, an average increase of 18-20 dB was observed in the signal-to-noise ratio of the denoised spectra. In the second stage of the study, Partial Least Squares Regression models were built using both the noisy and cleaned spectra and evaluated on the test set with root mean square error and coefficient of determination. The Partial Least Squares Regression models built with the denoised spectra achieved lower root mean square error and higher coefficient of determination values at all signal-to-noise ratio levels. Especially at the 10 dB signal-to-noise ratio level, the coefficient of determination value of the model increased from 0.44 to 0.71, while the root means square error decreased from 0.60 to 0.43. The results show that the deep learning-based AE architecture can effectively reduce random noise in near-infrared spectral data and significantly improve both spectral signal quality and quantitative modeling performance. This approach provides an effective solution to improve model reliability and accuracy in near-infrared spectroscopy analysis.

Keywords: autoencoder, milk analysis, deep learning, near-infrared spectroscopy, noise removal, PLSR.

✉Corresponding author. E-mail: *ozcancataltas@selcuk.edu.tr*

1. Introduction

In recent years, Near Infrared (NIR) spectroscopy has emerged as an important analytical method for rapid, non-destructive, and environmentally friendly analysis of chemical and physical properties in many industries such as agriculture, food, pharmaceuticals, and biotechnology (Cen & He, 2007; Pasquini, 2018). The main advantages of NIR spectroscopy include minimal sample preparation, the ability to analyze multiple components, and high throughput (Davies, 1993). However, the high dimensional and complex nature of NIR spectral data brings several challenges in the analysis process. In particular, random noise during measurement, interference from the instrument or environmental factors, and matrix effects can adversely affect the quality of the spectral data and, thus, the modeling accuracy (Rinnan et al., 2009; Roggo et al., 2007).

Various preprocessing techniques have traditionally been used to reduce the effect of noise in spectral data. Savitzky-Golay filtering, standard normal variation (SNV), derivatization, and orthogonal signal smoothing have been widely used to reduce systematic and random noise in spectral data (Barnes et al., 1989; Çataltaş & Tütüncü, 2021; Savitzky & Golay, 1964). However, these techniques can often lose a significant portion of the signal or fail to deal with complex noise structures. Furthermore, improper implementation of preprocessing steps can negatively affect the generalizability and reliability of the model (Engel et al., 2013).

In recent years, deep learning-based approaches have attracted attention in the literature for their potential in data analysis (Alexandre & Santos, 2024). In particular, autoencoders (AE) have been widely studied for their effectiveness in removing noise from spectral data. By learning low-dimensional representations of the data, AEs are able to reconstruct noise-free spectra (Lv et al., 2023). In particular, denoising AE architectures are successful in reconstructing the original signal by learning the noise artificially added to the input data. Compared to conventional preprocessing techniques, these methods are better able to model the complex and nonlinear structure of spectral data and provide more reliable noise-free data sets (Vincent et al., 2008). In recent years, there have been many studies on deep learning-based methods to improve noise removal or modeling performance in near-infrared spectroscopy. Lv et al. (2023) proposed a noise-tolerant approach for NIR quality monitoring applications with a stacked denoising autoencoder architecture and demonstrated higher accuracy than conventional methods. Zhang et al. (2019) reported significant achievements in both feature extraction and quantitative analysis in spectral data with DeepSpectra, an end-to-end deep learning-based model. Biancolillo and Marini (2018) compared various chemometric and machine learning-based methods for spectroscopic data in pharmaceutical analysis and highlighted the potential of deep learning approaches. Furthermore, Rinnan et al. (2009) presented an up-to-date review of data preprocessing and noise removal techniques in NIR spectroscopy, noting that deep learning-based methods show promising results, especially for complex and multidimensional data. These recent studies show that deep learning-based noise removal and modeling approaches in NIR spectroscopy offer significant advantages in terms of both accuracy and generalizability.

One of the most widely used methods for quantitative analysis of NIR spectra is Partial Least Squares Regression (PLSR). PLSR can effectively model the relationship between the target variable and the spectral matrix in high-dimensional and multivariate spectral data (Wold et al., 2001). However, noise in spectral data can negatively affect the accuracy and generalizability of PLSR models. In the literature, it has been shown that the performance of PLSR models can be significantly improved by removing noise from spectral data (Biancolillo & Marini, 2018; Zhang et al., 2019). Therefore, obtaining noise-free spectra is critical to improve the performance of PLSR and similar regression models.

In this study, different levels of artificial noise are added to NIR spectral data, and these noisy spectra are cleaned with a deep learning-based AE model. PLSR models are constructed using the denoised and noisy spectra, and their prediction performances are compared. Thus, the impact of AE-based noise removal on model accuracy and reliability in NIR spectroscopic analysis was systematically evaluated. The findings suggest that noise reduction in spectral data can be an important tool to improve model performance in quantitative analysis.

The main novelty of this study is the systematic evaluation of a deep learning-based denoising AE architecture for noise reduction in NIR spectral data and its impact on the performance of PLSR models. Unlike traditional preprocessing methods, the proposed approach leverages the power of deep learning to effectively remove complex and nonlinear noise, thereby improving both spectral signal quality and quantitative modeling accuracy.

The remainder of this paper is organized as follows: Section 2 describes the dataset and the methodology, including the noise addition process and the autoencoder architecture. Section 3 presents experimental results and discussion. Finally, Section 4 concludes the paper and suggests directions for future research.

2. Materials and methods

2.1. Dataset

In order to test the proposed method used in this study, an open-source NIR dataset was used (Diaz-Olivares et al., 2023). This dataset contains the spectrum of milk samples taken from different cows and at different times to determine the properties of cow's milk. Measurements were taken in the wavelength range from 960 nm to 1690 nm. The dataset contains different target parameter data, among which fat content was selected as the target parameter. The original spectra of the milk dataset are given in Figure 1.

This dataset was chosen because it offers a comprehensive and publicly available collection of NIR spectra from milk samples. It is very suitable for evaluating the robustness of noise removal techniques under real measurement conditions. In particular, the large number of samples makes it suitable for deep learning applications (Jiang et al., 2025). The dataset is suitable for testing the generalizability and effectiveness of the proposed denoising autoencoder approach, as it contains different sample diversity and measurement conditions. Furthermore, milk analysis is a common application of NIR spectroscopy, and improvements in this area could have important practical implications for the food industry (Fodor et al., 2024).

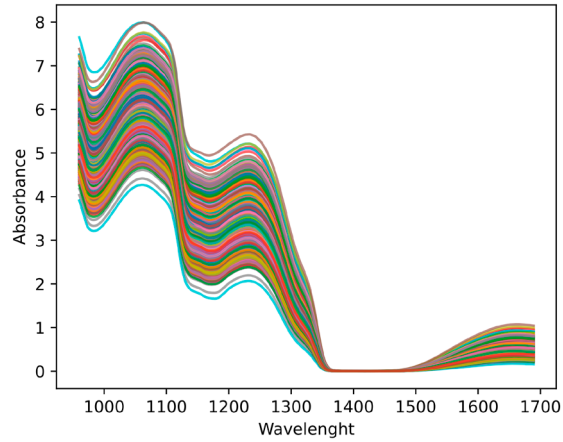

Figure 1. The original spectra of the milk dataset

The data set was divided into training and test subsets for preprocessing and modeling using random sampling. While 80% of the dataset was allocated as training data, 20% was allocated as test data to evaluate the generalizability of the model. In creating the training and test sets, the samples were randomly selected so that both sets represented the characteristics of the original dataset. This approach is widely used to prevent the model from overfitting during training and to provide a better fit to real-world data.

Each sample in the dataset has 255 spectral features representing measurements at different wavelengths between 960 nm and 1690 nm. In addition, each sample contains milk fat content information as a target parameter. Statistical information about the training and test subsets is given in Table 1.

Table 1. The descriptive statistics of the dataset analyzed

	Count	Average	Min	Max	Median	Std. Dev.
Train Subset	979	3.52	1.54	7.60	3.54	0.81
Test Subset	245	3.55	1.59	6.25	3.52	0.80
Dataset	1224	3.52	1.54	7.60	3.54	0.81

2.2. Artificial noise-adding process

Artificial Gaussian noise was added to the spectral data to simulate the random noise caused by real measurement conditions in NIR spectra and to evaluate the robustness of the model to noise. This approach has been used to study the effect of different signal-to-noise ratios (SNR) on model accuracy, to evaluate the performance of the denoising AE architecture, and to improve the generalizability of the model by providing data augmentation during training (Rinnan et al., 2009). The noise was modeled as white Gaussian noise and applied at four different SNR levels of 10, 15, 20, and 25 dB. These levels represent low, medium, and high noise intensity for milk and similar food spectra. Noise addition was performed in the following steps:

(1) Calculating the signal strength of the clean spectrum,

(2) Determining the noise power corresponding to the target SNR level,

(3) Generating random noise samples from a Gaussian distribution,

(4) Adding the generated noise to the clean spectrum.

Noise addition was applied separately to the training and test sets. A fixed initial value for the random number generator was used to ensure repeatability of the experiments. The accuracy of the noise addition process was checked by comparing the theoretical and measured SNR values and verified that the average deviation was less than 0.1 dB.

2.3. The autoencoder architecture

In this study, a denoising AE-based deep learning architecture is developed to remove noise from NIR spectra effectively. AE are artificial neural network models based on an unsupervised learning approach that encodes the input data into a lower dimensional representation and then attempts to reconstruct the original data from this representation. Denoising

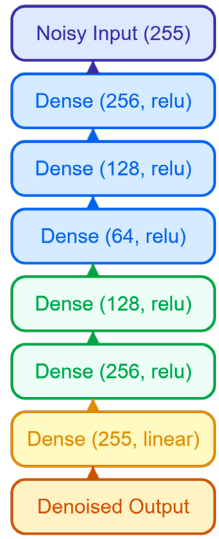

Figure 2. The model architecture of the designed autoencoder

autoencoders, on the other hand, aim to learn the noise artificially added to the input data and reconstruct the noise-reduced original signal. Thanks to these features, it can exhibit superior performance compared to classical preprocessing methods in removing complex and nonlinear noise structures in spectral data.

The autoencoder architecture used in this study consists of fully connected layers. The input layer of the model corresponds to the spectral data size. The encoder part contains three hidden layers consisting of 256, 128, and 64 neurons, respectively, and a nonlinear ReLU activation function is used in each layer. This structure allows the high dimensional and complex structure of the spectral data to be reduced to a more compact and meaningful representation. The decoder part consists of two layers of 128 and 256 neurons and is configured to achieve the original input size. In the output layer, a linear activation function is chosen so that the model can accurately reproduce the continuous values of the spectral data. The diagram of the designed AE model is given in Figure 2.

The autoencoder architecture presented in this paper is designed by the authors based on preliminary experiments and a review of recent literature. Consisting of three encoder and two decoder layers, it provides a good trade-off between model complexity and noise removal performance for NIR spectral data. The number of layers and neurons is experimentally determined to ensure efficient feature extraction and avoid overlearning. This architecture is not taken directly from any previous work but is explicitly tailored considering the characteristics of the milk NIR dataset and the requirements of the noise removal task.

During the training of the model, the mean squared error (MSE) is used as the loss function. The Adam algorithm was used for optimization, and the learning rate was set to 0.001. The maximum number of epochs was set to 500, and the batch size was set to 32. In order to

prevent overfitting, an early stopping strategy was applied by monitoring the validation loss. The patience value for early stopping was set to 10. The weights of the model were recorded at the epoch with the lowest validation loss. During the training process, a shuffle operation was applied to the training data, and a fixed seed was used to increase the generalizability of the model.

After the training was completed, the autoencoder model was used to predict the noisy spectra and denoised spectra were obtained. The performance of the model is evaluated with loss values calculated on both training and test data and visual spectrum comparisons. The noise removal performance is also quantitatively analyzed with theoretical and measured SNR values.

This architecture is in line with deep learning-based approaches proposed in the literature for noise reduction in NIR and other spectroscopic data and offers a more flexible and powerful solution compared to classical preprocessing techniques.

2.4. Partial least squares regression

In this study, Partial Least Squares Regression (PLSR) was used for quantitative estimation of chemical constituents of NIR spectral data. PLSR is a widely preferred regression technique for modeling the relationship between independent variables (spectra) and dependent variables (fat concentrations) in high-dimensional and multicollinear spectral data. By creating latent variables (LVs) that best explain the joint variance of the spectral matrix and the target variables, this method both reduces the data size and limits the risk of model overfitting (Wold et al., 2001). In building PLSR models, the PLSRegression class of the sci-kit-learn library was used (Michel et al., 2011).

2.5. Performance measures

Model performance was measured by root mean square error (RMSE) and coefficient of determination (R^2) metrics. RMSE indicates the magnitude of the model's prediction errors, while R^2 reflects the explanatory power of the model. By comparing the performance of models with noisy and cleaned spectra, the effect of autoencoder-based noise removal on PLSR model accuracy is quantitatively demonstrated. The mathematical formulation of RMSE and R^2 are given in Eqs. (1) and (2), respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n}\left(y_i - \widehat{y_i}\right)^2}\ ; \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y_i}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}. \tag{2}$$

In these formulas, y_i denotes the actual value of the target variable for the ith sample, $\widehat{y_i}$ is the predicted value for the ith sample, and n is the total number of samples in the test set. $\bar{y}$ represents the mean of the actual target values.

3. Results and discussion

In this study, different levels of artificial noise are added to the NIR spectra, and the noisy data are cleaned with a deep learning-based AE model. The noise removal performance is quantitatively evaluated both theoretically and quantitatively based on measured SNR values. Figure 3 shows the original, noisy, and denoised spectra obtained at different noise levels. The results are summarized in Table 2.

Table 2. The SNR values in noisy and denoised spectra

Target SNR	The mean SNR of noisy spectra	The mean SNR of denoised spectra	The difference in SNR values
SNR = 10	9.94 dB	29.62 dB	19.68 dB
SNR = 15	14.94 dB	34.16 dB	19.20 dB
SNR = 20	19.92 dB	37.96 dB	18.04 dB
SNR = 25	24.95 dB	39.45 dB	14.50 dB

The results show that the AE model provides a significant improvement at all SNR levels. The measured SNR values in the noisy spectra are approximately 10, 15, 20, and 25 dB, respectively, in line with the added artificial noise levels. After noise removal with AE, an average increase of more than 20 dB in SNR values was observed. Especially at the lowest SNR level (10 dB), the SNR value nearly tripled to 29.62 dB with AE. Similarly, at all other levels, the SNR values of the denoised spectrum increased significantly.

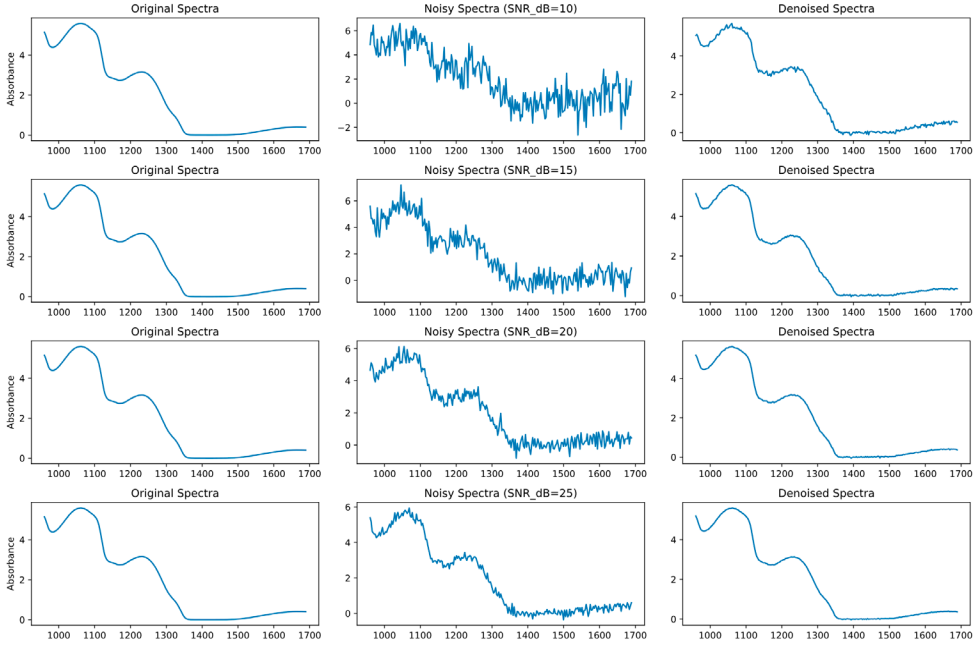

Figure 3. The original, noisy, and denoised spectra obtained at different noise levels

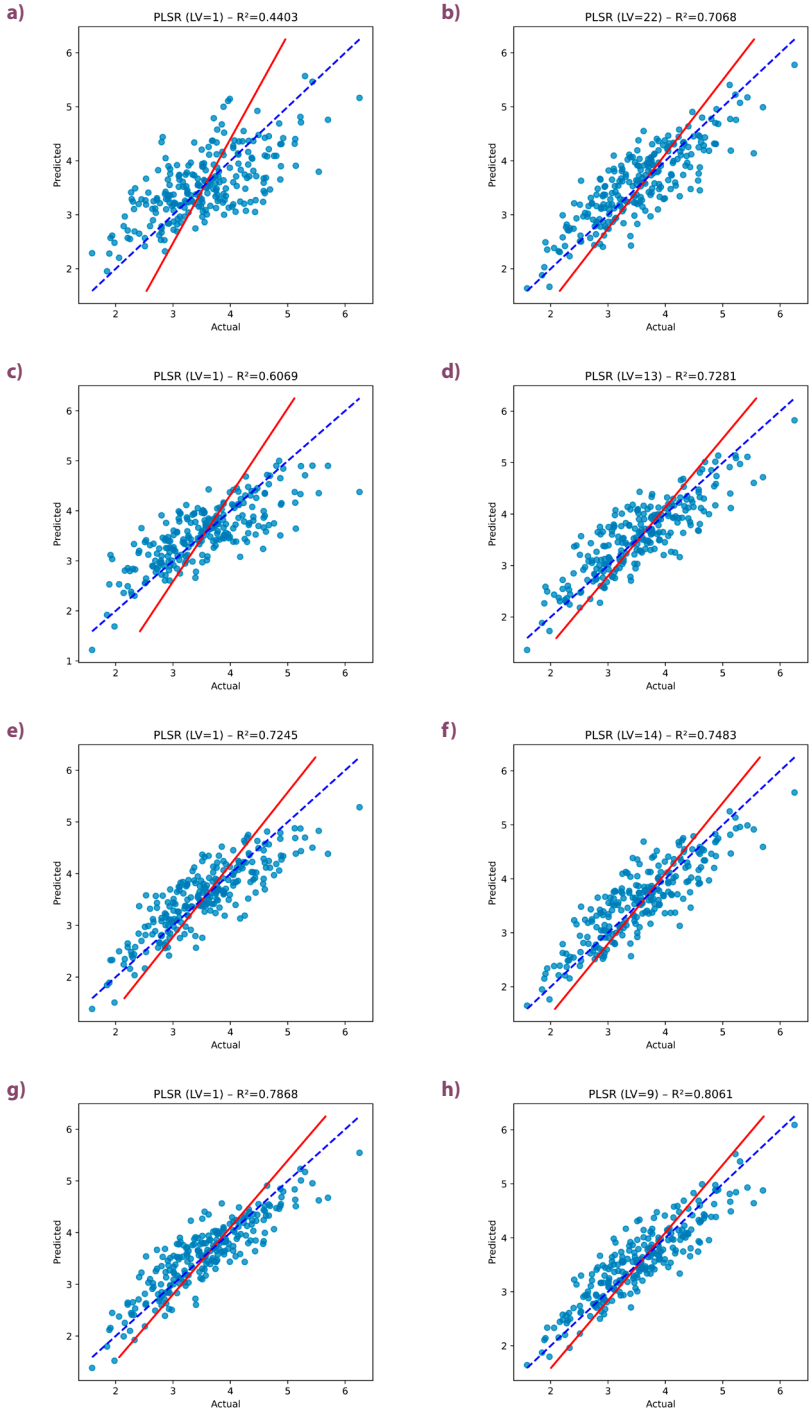

Figure 4. Comparison of PLSR Model Performances with different SNR values: a – Noisy-10 dB; b – Denoised-10 dB; c – Noisy-15 dB; d – Denoised-15 dB; e – Noisy-20 dB; f – Denoised-20 dB; g – Noisy-25 dB; h – Denoised-25 dB

In the modeling process, both noisy and autoencoder-noised spectra were evaluated separately. For both datasets, standard preprocessing techniques such as Savitzky-Golay (SG) filtering and SNV were applied prior to modeling. SG filtering reduces high-frequency noise in the spectral data, while SNV corrects for scattering and matrix effects. These preprocessing techniques are important to improve the accuracy and generalizability of the PLSR model. To control the complexity of the model and avoid overlearning, the optimal number of LVs was determined by 10-fold cross-validation. Cross-validation estimates were obtained for different numbers of LVs, and the root mean square error of cross-validation (RMSECV) was calculated. The number of LVs with the lowest RMSECV value was selected as the optimum for the final model.

The effect of noise removal on quantitative modeling is evaluated by the test set performance of PLSR models. The comparative results of R^2 values obtained with different SNR values in PLSR models are presented in Figure 4. The RMSE and R^2 values of the PLSR models using noisy and denoised spectra are presented in Table 3.

Table 3. Performance comparison of PLSR models with best LV using noisy and denoised spectra

Target SNR	Noisy			Denoised		
	LV	RMSE	R^2	LV	RMSE	R^2
SNR = 10	1	0.6	0.44	22	0.4343	0.7068
SNR = 15	1	0.5029	0.6069	13	0.4182	0.7281
SNR = 20	1	0.421	0.7245	14	0.4024	0.7483
SNR = 25	1	0.3703	0.7868	9	0.3531	0.8061

The results show that PLSR models built using spectra denoised with AE achieve lower RMSE and higher R^2 values than noisy spectra at all SNR levels. Especially at low SNR levels (SNR = 10 dB), the noise removal process led to a significant increase in model accuracy (R^2: from 0.44 to 0.71). Furthermore, the optimal number of latent variables increased after denoising, indicating that the cleaned spectra contain more structural information and that the model is able to learn more complex relationships.

The results clearly show that the denoising autoencoder significantly improves the quality of NIR spectral data, especially under low SNR conditions. The significant increase in SNR and R^2 values after denoising shows that the model is able to recover meaningful information that would otherwise be lost due to noise. This improvement is significant in practical applications where measurement noise is unavoidable. The increase in the optimal number of latent variables after noise removal indicates that the cleaned spectra contain more structural information, and the PLSR model is able to capture more complex relationships between the spectral data and the target variable. These findings are in line with recent studies in the literature and support the effectiveness of deep learning-based noise removal methods on spectral. Overall, the proposed approach not only improves prediction accuracy but also strengthens the robustness and reliability of quantitative models in real-world NIR spectroscopy applications.

4. Conclusions

In this study, a deep learning-based denoising autoencoder approach is proposed to enhance the robustness of near-infrared spectroscopy data against random noise and improve quantitative modeling accuracy. Different levels of artificial Gaussian noise are added to the NIR spectra obtained from milk samples, and these noisy data are efficiently cleaned by AE architecture. After noise removal, the signal-to-noise ratio of the spectral data is significantly improved by 18-20 dB on average. Especially under low SNR conditions, AE-based noise reduction improved the spectral signal quality by up to three orders of magnitude.

In the second phase of the study, Partial Least Squares Regression models were built using both noisy and AE-cleaned spectra, and model performance was evaluated on a test set. The results show that PLSR models with AE-noised spectra achieve lower RMSE and higher R^2 values at all SNR levels. Especially at the lowest SNR level, the R^2 value of the model increased from 0.44 to 0.71, while the RMSE decreased from 0.60 to 0.43. Moreover, it was observed that the optimum number of latent variables increased after denoising, and the model was able to learn more complex relationships.

The results show that the deep learning-based AE architecture can effectively reduce random noise in NIR spectral data and significantly improve both spectral signal quality and quantitative modeling performance. This approach provides an effective and innovative solution to improve model reliability and accuracy in NIR spectroscopic analysis. In the future, it would be helpful to investigate the effectiveness of AE-based noise removal approaches on different sample types and more complex spectral datasets in terms of generalizability and industrial applications.

References

Alexandre, R., & Santos, D. (2024). Just-In-Time Software Defect Prediction using a deep learning-based model. *New Trends in Computer Sciences*, *2*(2), 91–100. https://doi.org/10.3846/ntcs.2024.22274

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, *43*(5), 772–777. https://doi.org/10.1366/0003702894202201

Biancolillo, A., & Marini, F. (2018). Chemometric methods for spectroscopy-based pharmaceutical analysis. *Frontiers in Chemistry*, *6*, Article 412780. https://doi.org/10.3389/fchem.2018.00576

Çataltaş, Ö., & Tütüncü, K. (2021). A review of data analysis techniques used in near-infrared spectroscopy. *European Journal of Science and Technology*. https://doi.org/10.31590/ejosat.882749

Cen, H., & He, Y. (2007). Theory and application of near infrared reflectance spectroscopy in determination of food quality. *Trends in Food Science & Technology*, *18*(2), 72–83. https://doi.org/10.1016/j.tifs.2006.09.003

Davies, T. (1993). Book reviews: Practical NIR spectroscopy with applications in food and beverage analysis, Ft-NIR Atlas. *NIR News*, *4*(5), 12–12. https://doi.org/10.1255/nirn.212

Diaz-Olivares, J. A., van Nuenen, A., Gote, M. J., Díaz, V. F., Saeys, W., Adriaens, I., & Aernouts, B. (2023). Near-infrared spectra dataset of milk composition in transmittance mode. *Data in Brief*, *51*, Article 109767. https://doi.org/10.1016/j.dib.2023.109767

Engel, J., Gerretzen, J., Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M. C. (2013). Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, *50*, 96–106. https://doi.org/10.1016/j.trac.2013.04.015

Fodor, M., Matkovits, A., Benes, E. L., & Jókai, Z. (2024). The role of near-infrared spectroscopy in food quality assurance: A review of the past two decades. *Foods*, *13*(21), Article 3501. https://doi.org/10.3390/foods13213501

Jiang, Y., Ma, X., & Li, X. (2025). Towards virtual sample generation with various data conditions: A comprehensive review. *Information Fusion*, *117*, Article 102874. https://doi.org/10.1016/j.inffus.2024.102874

Lv, J., Chen, Z., Luan, X., & Liu, F. (2023). Denoising stacked autoencoders-based near-infrared quality monitoring method via robust samples evaluation. *The Canadian Journal of Chemical Engineering*, *101*(5), 2693–2703. https://doi.org/10.1002/cjce.24684

Michel, V., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pasquini, C. (2018). Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. *Analytica Chimica Acta*, *1026*, 8–36. https://doi.org/10.1016/j.aca.2018.04.004

Rinnan, Å., Berg, F. van den, & Engelsen, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, *28*(10), 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., & Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis*, *44*(3), 683–700. https://doi.org/10.1016/j.jpba.2007.03.023

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (*ICML '08*) (pp. 1096–1103). Association for Computing Machinery. https://doi.org/10.1145/1390156.1390294

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Zhang, X., Lin, T., Xu, J., Luo, X., & Ying, Y. (2019). DeepSpectra: An end-to-end deep learning approach for quantitative spectral analysis. *Analytica Chimica Acta*, *1058*, 48–57. https://doi.org/10.1016/j.aca.2019.01.002