



# Machine Learning-Enabled NIR Spectroscopy. Part 3: Hyperparameter by Design (HyD) Based ANN-MLP Optimization, Model Generalizability, and Model Transferability

Hussain Ali<sup>1</sup> · Prakash Muthudoss<sup>2,3,4</sup> · Chirag Chauhan<sup>5</sup> · Ilango Kaliappan<sup>6</sup> · Dinesh Kumar<sup>7</sup> · Amrit Paudel<sup>8,9</sup> · Gobi Ramasamy<sup>1</sup>

Received: 8 August 2023 / Accepted: 1 November 2023 / Published online: 7 December 2023  
© The Author(s) 2023

## Abstract

Data variations, library changes, and poorly tuned hyperparameters can cause failures in data-driven modelling. In such scenarios, model drift, a gradual shift in model performance, can lead to inaccurate predictions. Monitoring and mitigating drift are vital to maintain model effectiveness. USFDA and ICH regulate pharmaceutical variation with scientific risk-based approaches. In this study, the hyperparameter optimization for the Artificial Neural Network Multilayer Perceptron (ANN-MLP) was investigated using open-source data. The design of experiments (DoE) approach in combination with target drift prediction and statistical process control (SPC) was employed to achieve this objective. First, pre-screening and optimization DoEs were conducted on lab-scale data, serving as internal validation data, to identify the design space and control space. The regression performance metrics were carefully monitored to ensure the right set of hyperparameters was selected, optimizing the modelling time and storage requirements. Before extending the analysis to external validation data, a drift analysis on the target variable was performed. This aimed to determine if the external data fell within the studied range or required retraining of the model. Although a drift was observed, the external data remained well within the range of the internal validation data. Subsequently, trend analysis and process monitoring for the mean absolute error of the active content were conducted. The combined use of DoE, drift analysis, and SPC enabled trend analysis, ensuring that both current and external validation data met acceptance criteria. Out-of-specification and process control limits were determined, providing valuable insights into the model's performance and overall reliability. This comprehensive approach allowed for robust hyperparameter optimization and effective management of model lifecycle, crucial in achieving accurate and dependable predictions in various real-world applications.

**Keywords** Artificial Neural Network Multilayer Perceptron (ANN-MLP) · data-driven modelling · design of experiments (DoE) · hyperparameter optimization · model generalizability · model lifecycle management · model transferability · near infrared (NIR) · process monitoring · statistical process control (SPC) · target drift detection

## Highlights

Design of experiments (DoE): Efficient hyperparameter optimization of Artificial Neural Network Multilayer Perceptron (ANN-MLP) for enhanced model performance, generalizability, and transferability.  
Specification or threshold setting: Trend analysis and statistical process control (SPC) with business requirements (85–115%) aid in differentiating common-cause variation from special-cause variations, thereby mitigating their impact.  
Comprehensive approach: Amalgamation of Artificial Neural Network Multilayer Perceptron (ANN-MLP) with design of experiments (DoE), data drift analysis, and SPC leads to enhanced prediction and improved model lifecycle management (LCM).

## Abbreviations

AI/ML	Artificial intelligence/machine learning
ANN-MLP	Artificial Neural Network Multilayer Perceptron
API	Active pharmaceutical ingredient
BU	Blend uniformity
BVD	Bias-variance decomposition
CU	Content uniformity
CV	Cross-validation
DoE	Design of experiments
ICH	International Council for Harmonisation

Extended author information available on the last page of the article



K-S	Kolmogorov-Smirnov test
LCM	Life cycle management
MAE	Mean absolute error
MSE	Mean square error
MCR	Multivariate curve resolution
NIR/NIRS	Near-infrared spectroscopy
OOS	Out-of-specification
PLS	Partial least squares
R <sup>2</sup>	R-squared
RMSE	Root mean square error
SPC	Statistical process control
USFDA	United States Food and Drug Administration

## Introduction

Near-infrared spectroscopy (NIR) is a non-destructive analytical tool widely used in various industries [1]. It provides chemical and physical information by measuring the absorption of near-infrared light by a sample [1–3]. To decode this information, chemometrics, data pre-processing, and advanced analytical techniques such as artificial intelligence (AI) and machine learning (ML) deep learning are needed. NIR can be used in food [4–7], pharmaceutical [1–3, 8], alternative medicines like ayurveda [9], agricultural [10–13], dairy [14–16], and process analytics [17–19] to analyze raw materials, monitor processes, and evaluate final products without damaging them, thus offering a fast and cost-effective method of analysis.

Chemometrics methods such as multivariate curve resolution (MCR) and partial least squares (PLS) have been extensively used in the analysis of complex data sets [20–22]. However, in recent times, machine learning techniques such as multilayer perceptron or artificial neural networks (ANN) have gained popularity. Hussain *et al.* [23–25] pioneered the introduction of artificial neural networks in pharmaceutical science. Currently, it is extensively used owing to its capacity to manage large datasets and intricate variable relationships [3, 26, 27]. Hyperparameter tuning, model performance validation, and cross-validation are crucial steps in machine learning to ensure model accuracy and generalizability [28, 29]. Model interpretability [30], explainability [31], generalizability [3, 32], and transferability [3, 32] have also become essential factors in the development of machine learning models to facilitate their deployment in real-world applications. These topics are widely discussed in the literature, and their implementation can improve the reliability and robustness of machine learning models in various industries.

In our earlier works [2, 3, 33], we discussed methods for handling sampling, data pre-processing, data comprehension, model selection, model performance, model generalization, explainability, interpretability, and transferability.

Regarding open-source datasets, we found that a multilayer perceptron (MLP) model, trained using a lab-scale dataset, outperformed the partial least squares (PLS) model on pilot-scale data. However, the MLP model failed to deliver satisfactory results when applied to full/production-scale datasets [3]. It is therefore the need of the hour as well as a mandate to carefully evaluate the problems and options available to address such difficulties.

This research focuses on decoding the aforementioned issues and introduces a design of experiments-guided hyperparameter selection approach for ANN-MLP to achieve more reliable and consistent results [29]. By using design of experiments (DoE), the study systematically identifies the optimal combinations of neuron activation functions, hidden ANN layers, and max iterations, leading to the determination of the best ANN architecture. The DOE-based approach replaces conventional ANN-MLP training with a mathematically optimized exploration, making the ANN highly efficient for analyzing NIR data. Additionally, the implementation of DoE-guided ANN-MLP is anticipated to overcome the commonly encountered saddle-point challenges. The objectives of this study are as follows:

1. DoE of hyperparameters: optimize hyperparameters through design of experiments (DoE) to enhance model performance in the Artificial Neural Network Multilayer Perceptron (ANN-MLP)
2. Target drift detection: develop a structured methodology for detecting and addressing target drifts in the model's predictions
3. Specification criteria: determine the most suitable specification criteria, process/data shifts, and model transfer to production-scale
4. Trend analysis: implement trend analysis with a control plan and conduct failure mode effect analysis/root cause analysis using statistical process control (SPC) methods to ensure process reliability and quality control

## Materials and Methods

### Data Set and Data Understanding

In this investigation, a tablet dataset acquired employing near-infrared transmittance spectroscopy data related to Escitalopram® tablets (publicly available [34]) from the literature was utilized. Tablet data set was constructed through the analysis of 310 pharmaceutical tablets acquired employing NIR spectroscopy, which included around 400 wave numbers in the spectral wavenumber range between 7400 and 10,500 cm<sup>-1</sup>. The tablets were manufactured at the laboratory scale (lab scale), pilot scale, and production scale, for details refer [3, 34].

### Design-of-Experiments

The development of an D-optimal ANN-MLP model involved the application of design of experiments (DoE) and statistical modelling, which were performed using JMP Pro software version 14 by SAS Institute, NC, USA. ANN-DOE for multilayer perceptron constructed for predicting active content (in %w/w). The D-optimal DoE to determine ANN architecture (hidden layer, max iterations, activations, etc.) were carried out. The ANN model performance metrics considered were R2, MAE, MSE, RMSE, total time for modelling (in seconds), MSE for bias-variance decomposition, bias, variance, time taken for BVD (in seconds), and time taken for BVD (in minutes). All essential primary terms, second-level interaction terms, and curvilinear (quadratic terms) were included in the model to ensure comprehensive coverage of the variables and their potential interactions.

### Drift Monitoring

Evidently AI [35] is a tool that helps to monitor machine learning pipelines. It detects the following changes: (i) input feature distribution, hereinafter be termed as “Data Drift”; (ii) provides feature statistics and behavior overview; (iii) detect changes in dependent variable, hereinafter be termed as “Target Drift”; and (iv) evaluate the quality of machine learning model and errors, hereinafter be termed as “Model Drift” [36]. In this study, “Target Drift” is investigated, and model performance is evaluated accordingly. The researchers aim to understand how changes in the target variable over scale (lab, pilot, and full or production) impact the regression metrics and reliability of the model.

### Performance Metrics

The coefficient of determination ( $R^2$ ), mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE), which are specified in the following Eqs. 1–4, were each errors to assess the model’s predictive power [37, 38]. While the absolute values of the MAE, MSE, and RMSE results should be as low as feasible, the  $R^2$  ranges on a scale from 0 to 1 and should have higher values (>0.95).

$$R^2 = 1 - \left( \frac{\sum_{i=1}^n (y_i^{actual} - y_i^{pred})^2}{\sum_{i=1}^n (y_i^{actual} - y_i^{actual,mean})^2} \right) \tag{1}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i^{actual} - y_i^{pred}| \tag{2}$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i^{actual} - y_i^{pred})^2 \tag{3}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^{actual} - y_i^{pred})^2} \tag{4}$$

where  $y_i^{actual}$  and  $y_i^{pred}$  represented the reference and ML predicted values, respectively. On the other hand,  $y_i^{actual,mean}$  represents the experimental value meanwhile number of data points is represented as “n.”

### Training-Test Split

The training and test split of a dataset is necessary for evaluating the performance of algorithm and models’ predictability, generalizability, and transferability.

### Establishing Data-Split Criteria

#### Hold-out Dataset or Internal Validation or Model Generalizability

To choose the optimal ML model, the laboratory batch data used in this study served as the hold-out dataset. This data is divided using a random selection method in proportions of 80 (for the train): 20 (for the test). Using this hold-out approach, the effectiveness of the machine learning algorithms could be evaluated objectively.

#### External Validation Dataset or Model Transferability

The external validation sample could be made up of brand-new pilot or production-scale samples. This is precisely few of the ways how the model lifecycle management can be established. Additionally, a novel strategy known as internal–external validation architecture, is utilized by Muthudoss *et al.* [2]. This strategy combines the advantages of internal and external validation. The model performance on production scale batches is predicted using these two approaches.

### Bias-Variance Decomposition (BVD)

BVD used to analyze data and algorithm performance characteristics [2, 39]. Based on these results, the need for hyperparameter tuning was approached. The BVD demonstrates that mean squared error of a model generated by a certain algorithm is indeed made up of two components: (1) irreducible error (as noise) and (2) reducible error (as bias and variance), as shown in Eqs. 5 and 6. The irreducible error includes instrument/sample/sampling causes. On the other hand, bias measures how well predictions match the optimal values; variance indicates precision across different training sets that is considered crucial for evaluating model performance. Lowering bias and/or variance would allow in developing more accurate models. That said, a model with

minimal bias and minimal variance is better, often difficult to achieve. Hence, the bias-variance trade-off principle is employed,

$$\text{Mean squared error (model)} = \text{reducible error} + \text{irreducible error} \quad (5)$$

$$\text{Mean squared error (model)} = \text{bias}^2 (\text{model}) + \text{variance (model)} + \text{irreducible error} \quad (6)$$

## Data Analysis and Statistics

For this study, we explored the tablets datasets acquired using NIR as described in Ref [34] (<http://www.models.life.ku.dk/Tablets>). Python was used to analyze data using univariate and ML approaches (version 3.9.0). Machine learning models were built by using Scikit-learn package (version “0.24.0”) [40] and in Python. Evidently AI (version “0.24.0”) is used to monitor data/target/model drift [35]. Matplotlib package (version “3.4.1”) [41] was employed in generating plots. The bias-variance decomposition was performed using the mlxtend package (version “0.18.0”) [39] in Python (Raschka, 2018). Design of experiments (DoE) for ANN-MLP, statistical modelling, statistical analysis, and graphical visualization was carried out using JMP standard package (JMP®, Version 16, SAS Institute Inc. Cary, NC, 1989–2022).

## Results

### Design of Experiments and Hyperparameter Optimization

Classical design of experiments (like full factorial, fractional factorial, central composite, and Box-Behnken) involves manual planning and limited exploration. In contrast, computer-aided methods use algorithms and statistical methods to optimize designs efficiently, exploring factor spaces and identifying key factors affecting the response using minimal experiments. This approach enables informed decisions and better results with fewer experiments. A designed experiment is a controlled set of tests designed to model and explore the relationship between factors and one or more responses. With respect to ANN-MLP, the hyperparameters can be considered factors and the regression performance metrics can be considered responses. The regression performance metrics considered were R<sup>2</sup>, MAE, MSE, RMSE, total time for modelling (in seconds), MSE for bias-variance decomposition (BVD), bias, variance, time taken for BVD (in seconds), and time taken for BVD (in minutes). Two types of D-optimal (refer to Table I) design of experiments (DoE) were conducted to enhance the experimental design:

(i) a broader range of factors also termed “pre-screening DoE” (additionally identifies the optimal test size split) and (ii) a narrower range of factors also termed “optimization DoE” to optimize the pre-screened DoE. The first approach involved exploring a wide spectrum of factor levels to capture potential non-linear effects and interactions across a broader parameter space. On the other hand, the second approach focused on a more constrained range of factor levels, seeking to delve deeper into specific regions of interest and gain precise insights into the factors’ behavior within a narrower scope. By employing both strategies, the DoE aimed to comprehensively explore the effects of the factors and optimize the experimental design for more robust and informative results. A custom design platform in JMP software was utilized to create a 36-treatment pre-screening experimental design and 22-treatment optimization experimental design (Table I). This design allowed for estimates of main, interaction, and quadratic effects in predicting active content (in %w/w).

### Pre-screening DoE

The effect summary provides a concise overview of the significant effects observed in an experiment or statistical analysis, highlighting main factors, quadratic effects, and interactions with a statistically significant impact on the response variable. Researchers and decision-makers can use this summary to make informed decisions, optimize processes, and enhance the overall performance of the studied system. The analysis considers a *p*-value of less than 0.05 as significant, and examining the *p*-values for each factor helps determine their statistical significance. Additionally, understanding the nature of effects, whether linear or quadratic, further explains their influence on the response variable. Moreover, investigating interactions between factors reveals their combined impact on the response. The effect summary is used to screen model. Pre-screening DoEs in this study demonstrates the presence of main effects (hidden layer X, Y, Z, and max iterations) and all interactions between factors, as shown in Fig. 1a. Except max iterations, other factors did not demonstrate quadratic effect. Among the considered factors, pre-screening DOE indicated providing valuable insights for further analysis and decision-making.

### Prediction Profiler

The prediction profiler serves as a valuable tool for comprehending statistical model outcomes, enabling visualization of how predictor variables influence the response. Prediction profiles are especially useful in multiple-response models to help judge which factor values can optimize a complex set of criteria. Incorporation of bootstrapping through resampling enhances the analysis by generating

estimates of model parameters and evaluating prediction stability. Additionally, accounting for random noise factors in the process improves model performance, especially in complex scenarios. High  $R^2$  values, approaching 1, indicate a robust model fit, while low MAE, MSE, RMSE, time for modelling, MSE-BVD, bias, variance, time for BVD (in seconds), and time for BVD (in minutes) values signify precise predictions with minimal errors. This comprehensive approach empowers researchers to make informed decisions and enhance the reliability and effectiveness of statistical models across various real-world applications. By gaining deeper insights into the interplay between predictor variables and the response, researchers can better optimize their models for improved predictive performance. From effect summary, it was inferred that all the considered factors demonstrate main effects and interaction effects, while max iterations demonstrate quadratic effect as well. This is visible from the prediction profiler; except max iterations, all other factors show a linear dependency with the response while max iterations show a quadratic effect. In the design of experiments (DoE), a quadratic effect refers to the non-linear impact of a factor on the response variable, capturing curvature in the response surface, as shown in Fig. 1b.

### Optimization DoE

In the optimization design of experiments (DoEs), a fixed test size was employed, while other factors like hidden layer X, Y, Z, and max iterations were varied within a narrow range. To evaluate variability, bootstrapping with random noise was utilized. Most factors demonstrated linear dependencies, similar to the pre-screening DoE, except for max iterations. The max iterations factor exhibited a quadratic effect, indicating a non-linear relationship with the response. This quadratic effect revealed the presence of curvature in the response surface, indicating that changes in max iterations led to non-linear variations in the response variable. The findings from this analysis are crucial in understanding the optimal settings for the factors and identifying the

influence of each factor on the response variable, facilitating informed decision-making and process optimization, as shown in Fig. 2a below.

### Robustness of Optimized Model

Evaluating the robustness of a model involves testing its performance across a range of hyperparameter values, including lower, mid, and higher settings. By systematically varying the hyperparameters within their respective ranges, we can assess the model's ability to consistently deliver improved performance. If the model consistently performs better at mid-range or higher hyperparameter values, it indicates that these settings are optimal for achieving superior results. This observation suggests that the model is robust to variations in hyperparameter choices and can consistently perform well under different conditions. On the other hand, if the model shows poor or inconsistent performance across different hyperparameter values, it may indicate sensitivity to hyperparameter choices. In such cases, further experimentation and analysis may be needed to identify the best hyperparameter configuration for optimal performance. Understanding the robustness of the model by considering a range of hyperparameter values helps in selecting the most effective settings that lead to improved performance and reliability. Furthermore, this will help in overcoming the blackbox problems and the inconsistency in model performance.

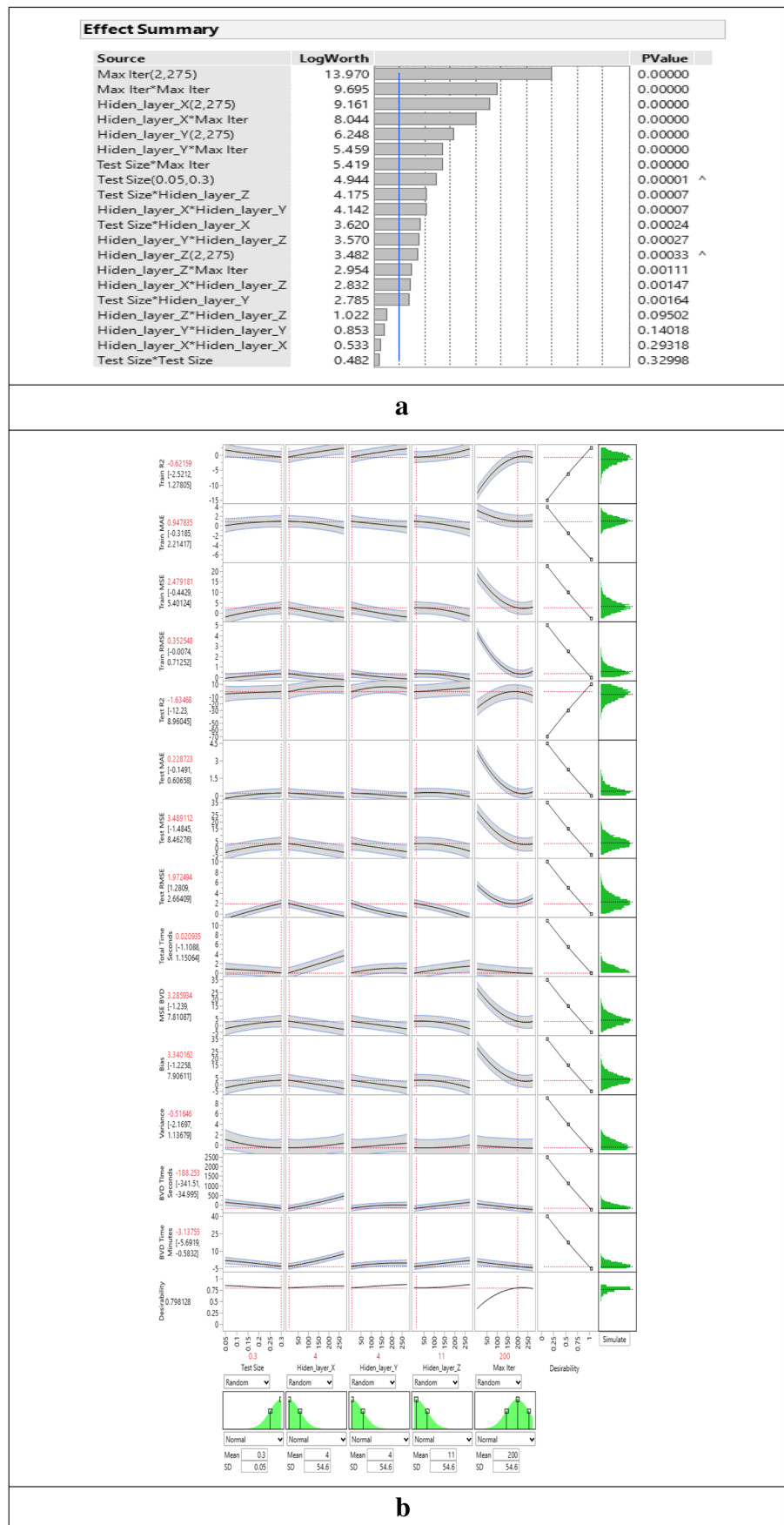
Table II demonstrates that the model's performance metrics, including  $R^2$ , MAE, MSE, RMSE, bias, variance, and time for model/BVD are acceptable for both training and test of the internal validation data. These results indicate that the model performs well and produces reliable as well as reproducible predictions. The metrics' values are satisfactory, signifying the model's effectiveness in capturing the underlying patterns in the data without overfitting or underfitting.

When the chosen hyperparameters consistently demonstrate exceptional performance across all relevant metrics, they are regarded as the optimal configuration for the model, as shown in Fig. 3. The fact that they consistently yield superior results in various aspects of the model's evaluation irrespective of the slight changes in the hyperparameters indicates their reliability and robustness. By excelling across multiple performance metrics, these hyperparameters ensure that the model performs consistently well in different scenarios and tasks. This reliability and versatility make the selected hyperparameters a preferred choice for achieving optimal and stable model performance. Researchers and practitioners can have confidence in their effectiveness and rely on them to deliver superior results in various real-world applications and data distributions.

**Table I** The Factors and Ranges Considered for the Two Different DoEs

Factors	Pre-screening DoE		Optimization DoE	
	Lower	Upper	Lower	Upper
Test size	0.05	0.3	Optimized and fixed to 0.3	
Hidden layer X	1	275	1	50
Hidden layer Y	1	275	1	50
Hidden layer Z	1	275	1	50
Max iterations	1	275	1	200

**Fig. 1 a** Effect summary for the DoE 1. **b** Prediction profiler for pre-screening DoEs



## Optimized Hyperparameters

The optimized hyperparameters that will be employed for external validation data are test size, 0.3; hidden layer X, 35; hidden layer Y, 4; hidden layer Y, 11; max iter, 179; activation = “identity”; learning\_rate = “invscaling”; solver = “lbfgs”; shuffle = false; and random\_state = 123.

## Target Drift Detection for the Dependent Features

Drift detection employing the K-S method can be utilized to understand whether the data come from similar or different distribution. If the data comes from the same distribution, the performance of the model can be described as model generalizability, while if the data comes from a different distribution and the model performs better, it is model transferability.

After optimizing the model architecture and hyperparameters using internal validation data, it becomes crucial to assess the similarity or dissimilarity between the external validation data and the data utilized during the optimization stage. This understanding is necessary to ensure the reliability and generalizability of the model. By comparing the characteristics and distribution of the external validation data with the data used for optimization, we can gain insights into potential variations, identify any potential drift, and evaluate how well the model performs on unseen data. It is possible to analyze the drift of independent features (NIR spectra), dependent features (active content in %w/w), or combination of both. Since influence of minor changes on NIR spectra is obvious, in this study, the drifts in dependent feature/active content is emphasized. The target drift report offered by evidently AI enables to delve into the modifications occurring in the target function and gain insights on how to adjust accordingly. Target or prediction drift signifies a scenario where the connection between the input variables (spectra) and the predicted target variable (API content in %w/w) undergoes changes over time. In simpler terms, the fundamental distribution of the target variable may shift, resulting in inaccurate or untrustworthy predictions. This drift can transpire due to alterations in the data generating process (NIR spectra), variations in sample characteristics, manufacturing process changes (lab to pilot to production, etc.), or shifts in the environment (humidity, temperature, etc.). From Table III, it can be inferred that no drifts in internal validation data (lab-scale train data vs. lab-scale test data) are observed. Similarly, no drift in external validation data (pilot-scale vs. full/production-scale) is observed. However, drifts in lab-scale vs. pilot-scale and lab-scale vs. full/production-scale are observed. In order to understand in depth the impact of such drifts, histogram was plotted, as shown in Fig. 4a–d.

Since the data points more or less follow the pattern for internal validation data (lab-scale train data vs. lab-scale test data) and external validation data (pilot-scale vs. full/production-scale), there is no drift observed. However, the lab-scale vs. pilot-scale or lab-scale vs. full/production-scale patterns does not match. Drift detected in the target drift algorithm could mean this. Since the data fall within the range of the trained model, the authors believe that the ANN-MLP algorithm and optimized hyperparameters should perform well.

In the context of the model, we can assess its generalizability by comparing the performance between lab scale train and lab scale test data since they originate from the same distribution. On the other hand, model transferability can be examined through (i) comparing lab scale train with pilot scale data and (ii) comparing lab scale train with production scale data, where the distribution of the target variable (active content in %w/w) differs, as summarized in Table IV. These comparisons allow us to determine how well the model can adapt and perform on data from different distributions or different scales of manufacture or different instruments or different excipients or domains, indicating its transferability to new and diverse NIR spectral data.

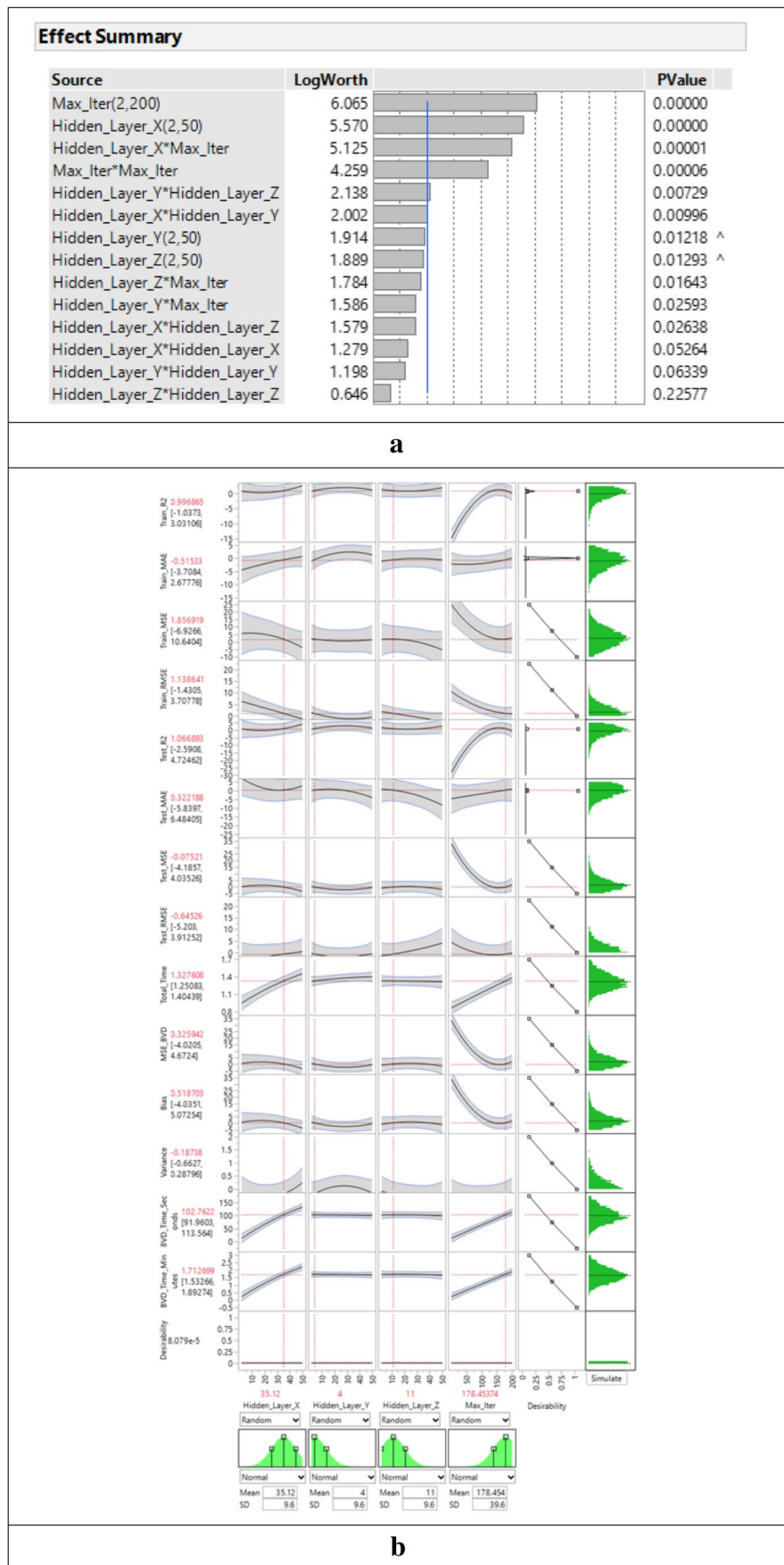
## Assessing Model Transferability

Before calculating point/interval estimate or employing control charts, it is crucial to assess whether there are any data drifts or distribution differences between the actual and predicted values, as well as the presence of outliers. This evaluation aids in determining whether the model requires retraining. Besides, this could be incorporated into model lifecycle management protocol for effective performance and reliability. Identifying drifts or discrepancies in the data helps in understanding potential shifts in the underlying patterns and the model’s reliability in current data conditions. By addressing these discrepancies through model retraining or adjustment, the model’s performance and effectiveness in predictions are ensured, leading to better decision-making based on reliable insights. Additionally, model transferability is also guaranteed. To this end, the drift detection method using the K-S test indicates  $p < 0.05$  for pilot-scale actual vs. predicted active content (refer Fig. 5a), showing drift, while full-scale actual vs. predicted has  $p > 0.05$ , implying no drift (refer Fig. 5b). This sensitivity of the target drift detection method suggests it can serve as a monitoring alarm. Consequently, precautionary trend analysis/control charts can be plotted to track the data proactively.

## Specification Limits and Trend Analysis

The test for uniformity of content in single-dose preparations involves assaying the individual contents of active substances in multiple units. It checks if each content is within

**Fig. 2 a** Effect summary for the optimization DoE. **b** Prediction profiler for optimization DoEs





**Table II** Performance Metrics of the Model (Robustness)

Metrics	Low	Mid	High
Train $R^2$	0.9466	0.9587	0.9582
Train MAE	0.1858	0.1730	0.1749
Train MSE	0.0711	0.0551	0.0556
Train RMSE	0.2666	0.2347	0.2359
Test $R^2$	0.9079	0.9336	0.9320
Test MAE	0.2658	0.2405	0.2450
Test MSE	0.2658	0.0746	0.0764
Test RMSE	0.2658	0.2731	0.2763
Time for modelling in seconds	1.1514	1.2972	1.6293
MSE-BVD	0.1130	0.0970	0.0990
Bias	0.0900	0.0760	0.0730
Variance	0.0230	0.0210	0.0260
Time for BVD in seconds	63.4700	103.2583	148.5502
Time for BVD in minutes	1.0578	1.7210	2.4758

85–115% of the average content. Conformance of a process to certain specification limits is one of the desired outcomes. The specification limit was achieved as mentioned in the paper (85 to 115%) which works out to be between 6.9 and 9.1% w/w. The internal validation data (Fig. 6a) includes both in-specification and out-of-specification (OOS) data, while the external validation data (Fig. 6b) complies with the specified requirements.

### Control Chart and Trend Analysis

Typically, control charts and/or trend analysis act as monitoring thresholds for numerical decision-making purposes [42–44]. The paper employs trend analysis and control charts to assess whether the predictions generated by the ANN model fall within the specified limits. These visualizations are used to monitor both the internal validation data (lab scale train and lab scale test) and external validation data (pilot scale and full/production scale), ensuring the consistency and performance of the process. By utilizing these tools, the study demonstrates

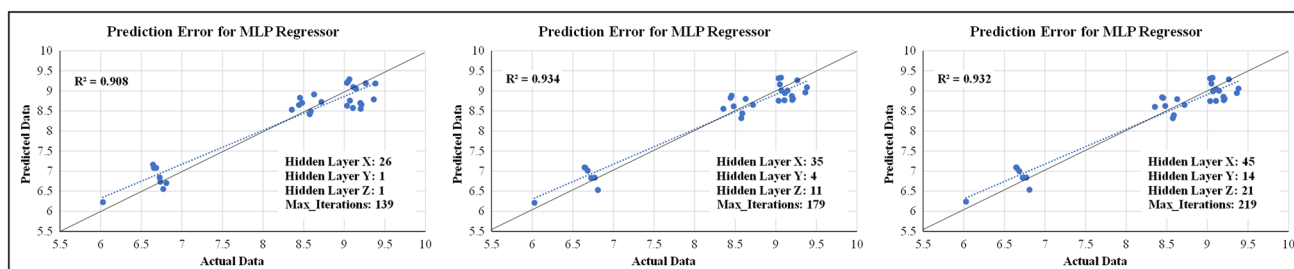
**Table III** Evaluating the Target (Active Content in %w/w) Drift as a Function of Manufacturing Process

	Drift detection	Detection method	Drift score ( $p$ -value)
Lab (train vs. test)	Not detected	K-S	0.166
Lab (train vs. pilot)	Detected	K-S	0.00
Lab (train vs. full)	Detected	K-S	0.00
Pilot vs. full	Not detected	K-S	0.22

the effectiveness of the process in adhering to the specified limits and identifies any potential deviations from the desired outcomes. Control charts play a crucial role in detecting trends, patterns, and outliers, enabling timely adjustments or interventions to maintain process quality. The relative error rate, computed using the lab scale train data (actual or original), is illustrated in Fig. 7. The ANN model accurately predicts out-of-specification results for lab scale train and test data, indicating its capability to identify deviations. Furthermore, it successfully predicts within specification limits for pilot scale and full-scale data, showcasing its robustness, generalizability, and transferability across different scales and datasets.

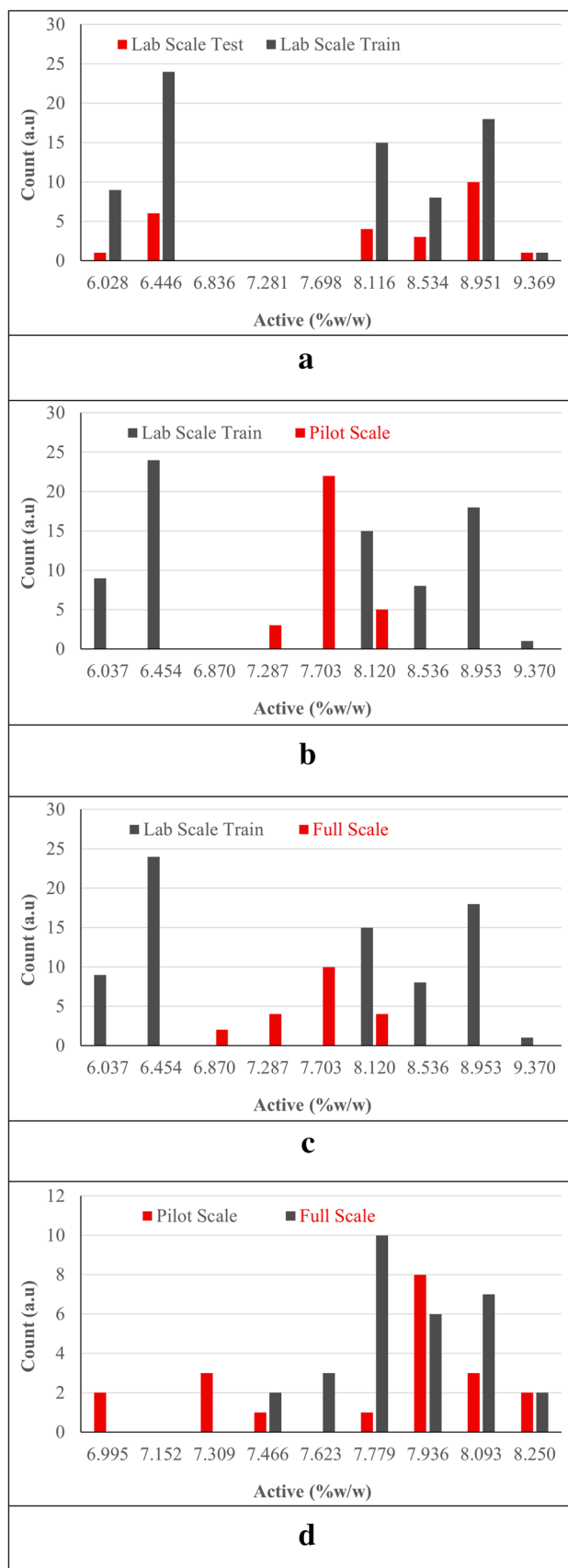
### Discussion

With respect to pharmaceutical products and processes, both the US Food and Drug Administration (USFDA) and the International Council for Harmonisation of Technical Requirements for Human Use (ICH) have established regulatory guidelines to address variation [45–48]. The primary objective of these guidelines is to exert control over variability and promote the adoption of scientific and risk-based approaches for assessing, managing, and controlling variation in pharmaceutical processes [49–51]. This study specifically focuses on addressing variability in Artificial Neural Network Multilayer Perceptron (ANN-MLP) models



**Fig. 3** Validation of the optimization DoE

**Fig. 4** **a** Target drift detection for internal validation data lab scale (train vs. test). **b** Target drift detection for external validation data (lab scale train vs. pilot scale). **c** Target drift detection for external validation data lab scale (train vs. full/production scale). **d** Target drift detection for external validation data (pilot scale vs. full/production scale)



**Table IV** Demonstrates the Evaluation of Distribution Similarity to Determine the Model’s Generalizability and Transferability on Unseen Data

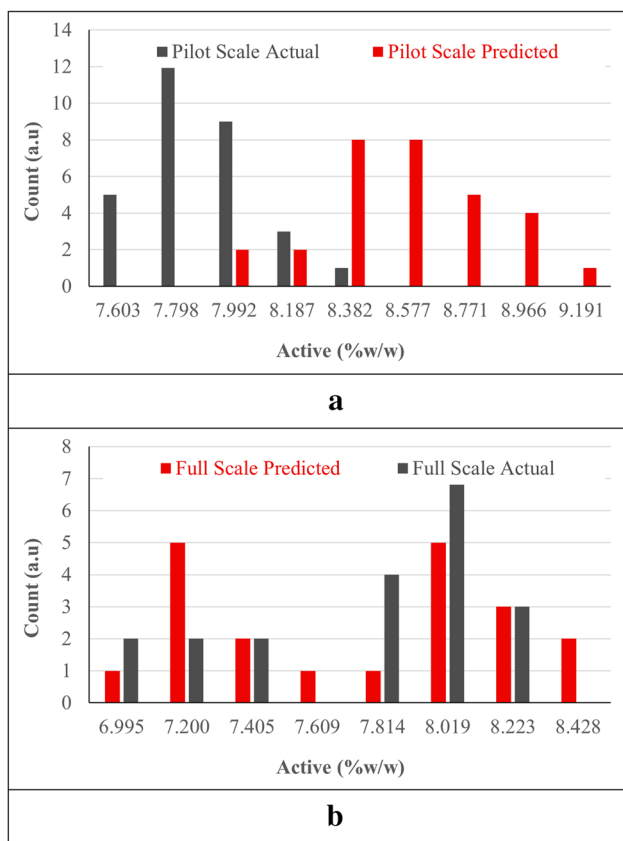
	Distribution-same	Distribution-different
ML prediction (successful)	Model generalizability	Model transferability
In this study	Internal validation data (lab scale train vs. lab scale test)	External validation data (lab scale train vs. pilot scale and lab scale train vs. production scale)

(especially hyperparameter tuning), which are powerful tools used in various applications, including pharmaceutical research and development. To meet the regulatory requirements and ensure product quality, safety, and efficacy, a systematic approach is employed.

The study utilizes design of experiments (DoE) to identify sources of variability in ANN-MLP models. This approach allows researchers to comprehensively understand the factors that influence the model’s performance. By identifying these factors, researchers can take measures to control and

optimize the model’s behavior. To maintain control over the ANN-MLP model, a target drift or a Kolmogorov–Smirnov test is employed. This ensures that the model remains within desired limits and avoids any potential issues related to out-of-trend or out-of-specification results. By closely monitoring the model’s performance, researchers can take timely corrective actions if any deviations are detected, preventing potential quality problems. Variability in the ANN-MLP model is further monitored through trend analysis and statistical process control (SPC). This allows researchers to evaluate the model’s performance over time and identify any deviations from the expected behavior. Early detection of variations enables researchers to fine-tune the model and ensure its stability and reliability.

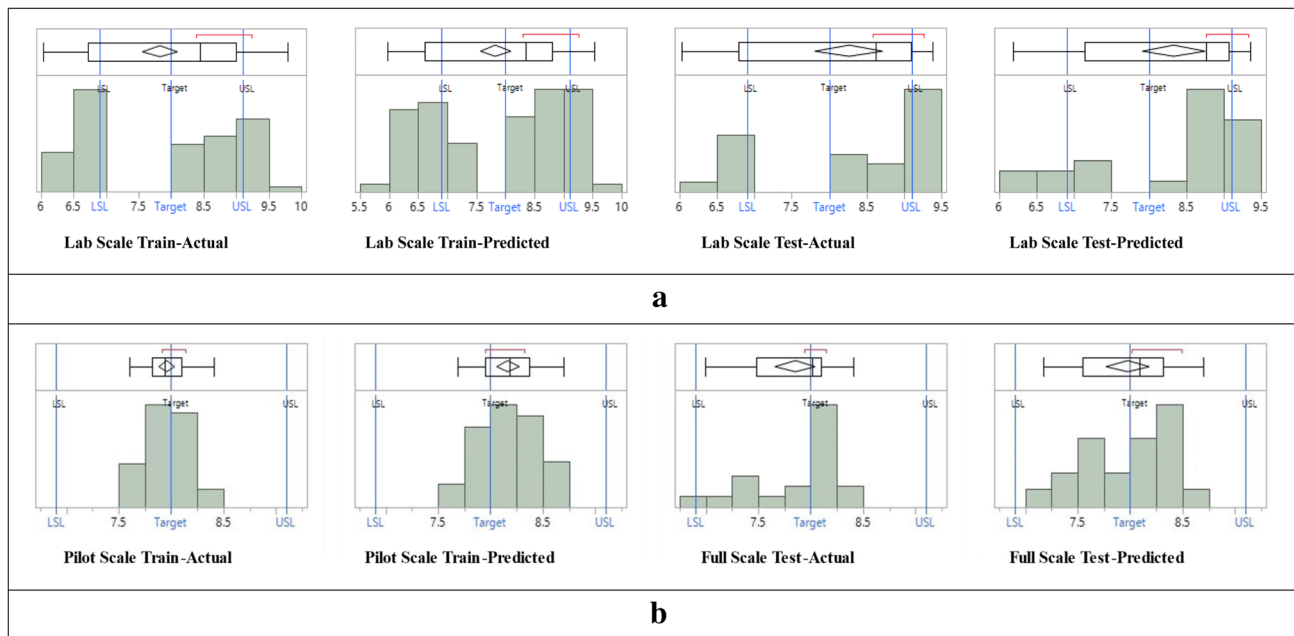
By integrating these methodologies, the study aims to enhance the stability and reliability of ANN-MLP models. This makes them more adaptable to diverse datasets and real-world applications, providing robust and reliable tools for various pharmaceutical processes. The ultimate goal is to ensure compliance with USFDA guidelines and promote a scientific and risk-based approach to achieve optimal product quality while mitigating potential risks associated with variability in pharmaceutical processes.



**Fig. 5** a Target drift detection for pilot scale actual vs. predicted. b Target drift detection for full scale actual vs. predicted

### Conclusions

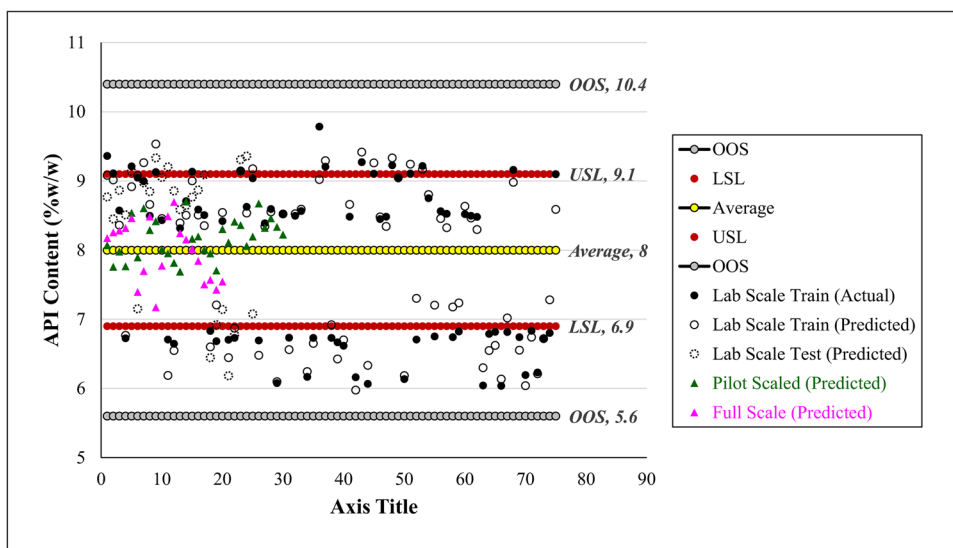
In conclusion, this study underscores the importance of addressing ML model drift or errors arising from data variations, library changes, and missing parameter information to ensure accurate predictions. The comprehensive approach integrating DoE, drift analysis, and SPC enables robust hyperparameter optimization and effective model lifecycle management, leading to dependable predictions in real-world scenarios. Continuous monitoring and mitigation of model drift are essential to maintain model effectiveness over time. The findings contribute to enhancing decision-making processes and optimizing model performance in dynamic data environments. Overall, this study provides valuable insights into managing and achieving reliable predictive models that can be integrated in production environments.



**Fig. 6 a** Specification Limit (85%–115% w/w) and Internal Validation Data. Note: The specification limit is between 6.9% w/w and 9.1% w/w. The internal validation data (Lab Scale Train Actual vs. Predicted and Lab Scale Test Actual vs. Predicted) includes both in-specification and out-of-specification (OOS) data. **b** Specification

Limit (85%–115% w/w) and External Validation Data. Note: The specification limit is between 6.9% w/w and 9.1% w/w. The external validation data (Pilot Scale Actual vs. Predicted and Full Scale Actual vs. Predicted) includes both in-specification and out-of-specification (OOS) data

**Fig. 7** Individual data points demonstrating the predicted active content %w/w



**Acknowledgements** HA and PM acknowledge AI and machine learning interactions with Saurabh Shahane, The Machine Learning Company (<https://themlco.com/>), Mumbai, India, and Vetrivel PS, Accenture, Chennai, India, and blog (<https://thehackweekly.com/>).

**Funding** Open access funding provided by Graz University of Technology.

**Data Availability** The data that support the findings of this study are available from the corresponding author, Prof. Dr. Amrit Paudel, upon reasonable request.

## Declarations

**Conflict of Interest** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.





## References

- Saravanan D, Muthudoss P, Khullar P, Rose VA. Quantitative microscopy: particle size/shape characterization, addressing common errors using 'analytics continuum' approach. *J Pharm Sci Elsevier*. 2021;110:833–49.
- Muthudoss P, Tewari I, Chi RLR, Young KJ, Ann EYC, Hui DNS, et al. Machine learning-enabled NIR spectroscopy in assessing powder blend uniformity: clear-up disparities and biases induced by physical artefacts. *AAPS PharmSciTech*. 2022;23:277 (Springer).
- Ali H, Muthudoss P, Ramalingam M, Kanakaraj L, Paudel A, Ramasamy G. Machine learning-enabled NIR spectroscopy. Part 2: workflow for selecting a subset of samples from publicly accessible data. *AAPS PharmSciTech*. 2023;24:34 (Springer).
- Cayuela-Sánchez JA, Palarea-Albaladejo J, García-Martín JF, del Carmen Pérez-Camino M. Olive oil nutritional labeling by using Vis/NIR spectroscopy and compositional statistical methods. *Innov Food Sci Emerg Technol*. 2019;51:139–47 (Elsevier).
- de Oliveira Moreira AC, Braga JWB. Authenticity identification of copaiba oil using a handheld NIR spectrometer and DD-SIMCA. *Food Anal Methods Springer*. 2021;14:865–72.
- Mauer LJ, Taylor LS. Water-solids interactions: deliquescence. *Annual review of food science and technology*. 2010;10:1:41–63.
- Kar S, Tudu B, Jana A, Bandyopadhyay R. FT-NIR spectroscopy coupled with multivariate analysis for detection of starch adulteration in turmeric powder. *Food Addit Contam Part A*. 2019;36:863–75 (Taylor & Francis).
- Saravanan D, Muthudoss P, Khullar P, Rosevenis A. Vendor qualification: utilization of solid state characterization "Toolbox" to assess material variability for active pharmaceutical ingredient. *J Appl Pharm Sci*. 2019;9:1–9.
- Rajesh PKS, Kumaravelu C, Gopal A, Suganthi S. Studies on identification of medicinal plant variety based on NIR spectroscopy using plant leaves. 2013 15th Int Conf Adv Comput Technol. 2013. p. 1–4.
- Mishra P, Herrmann I, Angileri M. Improved prediction of potassium and nitrogen in dried bell pepper leaves with visible and near-infrared spectroscopy utilising wavelength selection techniques. *Talanta*. 2021;225:121971 (Elsevier).
- Mishra P, Roger JM, Marini F, Biancolillo A, Rutledge DN. Parallel pre-processing through orthogonalization (PORTO) and its application to near-infrared spectroscopy. *Chemom Intell Lab Syst*. 2021;212:104190 (Elsevier).
- Mishra P, Roger JM, Rutledge DN, Woltering E. SPORT pre-processing can improve near-infrared quality prediction models for fresh fruits and agro-materials. *Postharvest Biol Technol*. 2020;168:111271 (Elsevier).
- Sampaio PS, Brites CM. Near-Infrared spectroscopy and machine learning: analysis and classification methods of rice. *Integrative Advances in Rice Research*. 2022;26:257. <https://doi.org/10.5772/intechopen.99017>.
- Pi F, Shinzawa H, Ozaki Y, Han D. Non-destructive determination of components in processed cheese slice wrapped with a polyethylene film using near-infrared spectroscopy and chemometrics. *Int Dairy J*. 2009;19:624–9 (Elsevier).
- Wang Y, Ding W, Kou L, Li L, Wang C, Jurick WM. A non-destructive method to assess freshness of raw bovine milk using FT-NIR spectroscopy. *J Food Sci Technol*. 2015;52:5305–10 (Springer).
- Núñez-Sánchez N, Martínez-Marín AL, Polvillo O, Fernández-Cabanás VM, Carrizosa J, Urrutia B, et al. Near infrared spectroscopy (NIRS) for the determination of the milk fat fatty acid profile of goats. *Food Chem*. 2016;190:244–52 (Elsevier).
- Rish AJ, Henson SR, Alam A, Liu Y, Drennen JK, Anderson CA. Comparison between pure component modeling approaches for monitoring pharmaceutical powder blends with near - infrared spectroscopy in continuous manufacturing schemes. *AAPS J [Internet]*. Springer International Publishing; 2022;24:1–10. <https://doi.org/10.1208/s12248-022-00725-x>
- Sulub Y, Wabuye B, Gargiulo P, Pazdan J, Cheney J, Berry J, et al. Real-time on-line blend uniformity monitoring using near-infrared reflectance spectrometry: a noninvasive off-line calibration approach. *J Pharm Biomed Anal*. 2009;49:48–54.
- Ni W, Nørgaard L, Mørup M. Non-linear calibration models for near infrared spectroscopy. *Anal Chim Acta [Internet]*. Elsevier B.V.; 2014;813:1–14. <https://doi.org/10.1016/j.aca.2013.12.002>
- Mishra P, Nordon A, Roger J-M. Improved prediction of tablet properties with near-infrared spectroscopy by a fusion of scatter correction techniques. *J Pharm Biomed Anal*. 2021;192:113684 (Elsevier).
- Kessler W, Oelkrug D, Kessler R. Using scattering and absorption spectra as MCR-hard model constraints for diffuse reflectance measurements of tablets. *Anal Chim Acta*. 2009;642:127–34 (Elsevier).
- Rebiere H, Ghyselinck C, Lempereur L, Brenier C. Investigation of the composition of anabolic tablets using near infrared spectroscopy and Raman chemical imaging. *Drug Test Anal*. 2016;8:370–7 (Wiley Online Library).
- Hussain AS, Shivanand P, Johnson RD. Application of neural computing in pharmaceutical product development: computer aided formulation design. *Drug Dev Ind Pharm*. 1994;20:1739–52.
- Dowell JA, Hussain A, Devane J, Young D. Artificial neural networks applied to the in vitro-in vivo correlation of an extended-release formulation: initial trials and experience. *J Pharm Sci*. 1999;88:154–60.
- Hussain AS, Yu X, Johnson RD. Application of neural computing in pharmaceutical product development. *Pharm Res*. 1991;8:1248–52 (Springer).
- Sarker IH. Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci*. 2021;2:160 (Springer).
- Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, Zhao S. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* 2019;18(6):463–77.

28. Arboretti R, Ceccato R, Pegoraro L, Salmaso L. Design of experiments and machine learning for product innovation : a systematic literature review. *Qual Reliab Eng Int.* 2022;38:1131–56.
29. Rodriguez-Granrose D, Jones A, Loftus H, Tandeski T, Heaton W, Foley KT, et al. Design of experiment (DOE) applied to artificial neural network architecture enables rapid bioprocess improvement. *Bioprocess Biosyst Eng.* 2021;44:1301–8 (Springer).
30. Gaurav D, Tiwari S. Interpretability vs explainability: the black box of machine learning. *2023 Int Conf Comput Sci Inf Technol Eng.* 2023. p. 523–8.
31. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable ai: a review of machine learning interpretability methods. *Entropy.* 2020;23:18 (MDPI).
32. Albahra S, Gorbett T, Robertson S, D'Aleo G, Kumar SV, Ockunzzi S, Lallo D, Hu B, Rashidi HH. Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts. *Semin Diagn Pathol* 2023;40(2):71–87.
33. Chaudhary S, Muthudoss P, Madheswaran T, Paudel A, Gaikwad V. Artificial intelligence (AI) in drug product designing, development, and manufacturing. In: *A handbook of artificial intelligence in drug delivery.* Academic Press; 2023. p. 395–442.
34. Dyrby M, Engelsen SB, Nørgaard L, Bruhn M, Lundsberg-Nielsen L. Chemometric quantitation of the active substance (containing C≡N) in a pharmaceutical tablet using near-infrared (NIR) transmittance and NIR FT-Raman spectra. *Appl Spectrosc.* 2002;56:579–85.
35. <https://www.evidentlyai.com/>. Open-source machine learning monitoring. Accessed 14th Nov 2023.
36. Madkour AH, Elsayed A, Abdel-Kader H. Historical Isolated Forest for detecting and adaptation concept drifts in nonstationary data streaming. *IJCI. Int J Comput Inf.* 2023;10(2):16–27.
37. Andersen CM, Bro R. Variable selection in regression—a tutorial. *J Chemom.* 2010;24:728–37 (Wiley Online Library).
38. Rajalahti T, Kvalheim OM. Multivariate data analysis in pharmaceuticals: a tutorial review. *Int J Pharm.* 2011;417:280–90 (Elsevier).
39. Raschka S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *J Open Source Softw.* 2018;3:638 (The Open Journal).
40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825–30 (JMLR.org).
41. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng.* 2007;9:90–5 (IEEE Computer Society).
42. Lörchner C, Horn M, Berger F, Fahl-Hassek C, Glomb MA, Esslinger S. Quality control of spectroscopic data in non-targeted analysis—development of a multivariate control chart. *Food Control.* 2022;133:108601 (Elsevier).
43. Malindzakova M, Čulková K, Trpčevská J. Shewhart control charts implementation for quality and production management. *Processes MDPI.* 2023;11:1246.
44. Pérez-Benítez BE, Tercero-Gómez JG, Khakifirooz M. A review on statistical process control in healthcare: data-driven monitoring schemes. *IEEE Access.* 2023;11:56248–272. <https://doi.org/10.1109/ACCESS.2023.3282569> (<https://ieeexplore.ieee.org/abstract/document/10144935>).
45. Yu LX. Pharmaceutical quality by design: product and process development, understanding, and control. *Pharm Res.* 2008;25:781–91.
46. Pluta PL. FDA lifecycle approach to process validation—what, why, and how? *J Valid Technol.* 2011;17:51 (MultiMedia Healthcare Inc).
47. U.S. Food and Drug Administration (USFDA), Process validation: general principles and practices, ID: FDA-2008-D-0559, USFDA, Silver Spring, MD 2011.
48. Lange R, Schnor T. Product quality, quality control and validation. In: *Practical Pharmaceutics: an international guideline for the preparation, care and use of medicinal products.* Cham: Springer International Publishing; 2023. p. 767–83.
49. Pramod K, Tahir MA, Charoo NA, Ansari SH, Ali J. Pharmaceutical product development: a quality by design approach. *Int J Pharm Investig.* 2016;6:129 (Wolters Kluwer--Medknow Publications).
50. Kovács B, Kovács-Deák B, Székely-Szentmiklósi I, Fülöp I, Bába L-I, Boda F, et al. Quality-by-design in pharmaceutical development: from current perspectives to practical applications. *Acta Pharm.* 2021;71:497–526 (Hrvatsko farmaceutsko društvo).
51. Lee S-H, Kim J-K, Jee J-P, Jang D-J, Park Y-J, Kim J-E. Quality by design (QbD) application for the pharmaceutical development process. *J Pharm Investig.* 2022;52:649–82 (Springer).

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

Hussain Ali<sup>1</sup> · Prakash Muthudoss<sup>2,3,4</sup>  · Chirag Chauhan<sup>5</sup> · Ilango Kaliappan<sup>6</sup> · Dinesh Kumar<sup>7</sup>  · Amrit Paudel<sup>8,9</sup>  · Gobi Ramasamy<sup>1</sup> 

✉ Amrit Paudel  
amrit.paudel@tugraz.at

✉ Gobi Ramasamy  
gobi.r@christuniversity.in

<sup>1</sup> Christ (Deemed to Be University), Bangalore 560029, Karnataka, India

<sup>2</sup> A2Z4.0 Research and Analytics Private Limited, Chennai 600062, Tamilnadu, India

<sup>3</sup> NuAxon Bioscience Inc., Bloomington, Indiana 47401-6301, USA

<sup>4</sup> School of Pharmaceutical Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Velan Nagar P.V. Vaithiyalingam Road Pallavaram 600117, Chennai, Tamilnadu, India

<sup>5</sup> The Machine Learning Company, Pune, India

<sup>6</sup> School of Pharmacy, Hindustan Institute of Technology and Science (HITS), Padur, 603 103 Chennai, Tamilnadu, India

<sup>7</sup> Department of Pharmaceutical Engineering & Technology, IIT (BHU), Varanasi 221011, Uttar Pradesh, India

<sup>8</sup> Research Center Pharmaceutical Engineering GmbH (RCPE), Inffeldgasse 13, 8010 Graz, Austria

<sup>9</sup> Graz University of Technology, Institute of Process and Particle Engineering, Inffeldgasse 13/3, 8010 Graz, Austria