AAPS PharmSciTech (2023) 24:34
https://doi.org/10.1208/s12249-022-02493-5

RESEARCH ARTICLE

Applications of Machine Learning and A.I. in Pharmaceutical Development and Technology

Machine Learning–Enabled NIR Spectroscopy. Part 2: Workflow for Selecting a Subset of Samples from Publicly Accessible Data

Hussain Ali[1] · Prakash Muthudoss[2] · Manikandan Ramalingam[3] · Lakshmi Kanakaraj[3] · Amrit Paudel[4,5] · Gobi Ramasamy[1]

Received: 20 October 2022 / Accepted: 14 December 2022 / Published online: 10 January 2023

Abstract

An increasingly large dataset of pharmaceutics disciplines is frequently challenging to comprehend. Since machine learning needs high-quality data sets, the open-source dataset can be a place to start. This work presents a systematic method to choose representative subsamples from the existing research, along with an extensive set of quality measures and a visualization strategy. The preceding article (Muthudoss *et al.*. in AAPS PharmSciTech 23, 2022) describes a workflow for leveraging near infrared (NIR) spectroscopy to obtain reliable and robust data on pharmaceutical samples. This study describes the systematic and structured procedure for selecting subsamples from the historical data. We offer a wide range of in-depth quality measures, diagnostic tools, and visualization techniques. A real-world, well-researched NIR dataset was employed to demonstrate this approach. This open-source tablet dataset (http://www.models.life.ku.dk/Tablets) consists of different doses in milligrams, different shapes, and sizes of dosage forms, slots in tablets, three different manufacturing scales (lab, pilot, production), coating differences (coated *vs* uncoated), etc. This sample is appropriate; that is, the model was developed on one scale (in this research, the lab scale), and it can be great to investigate how well the top models are transferable when tested on new data like pilot-scale or production (full) scale. A literature review indicated that the PLS regression models outperform artificial neural network-multilayer perceptron (ANN-MLP). This work demonstrates the selection of appropriate hyperparameters and their impact on ANN-MLP model performance. The hyperparameter tuning approaches and performance with available references are discussed for the data under investigation. Model extension from lab-scale to pilot-scale/production scale is demonstrated.

Highlights

- We present a comprehensive quality metrics and visualization strategy in selecting subsamples from the existing studies
- A comprehensive assessment and workflow are demonstrated using historical real-world near-infrared (NIR) data sets
- Selection of appropriate hyperparameters and their impact on artificial neural network-multilayer perceptron (ANN-MLP) model performance
- The choice of hyperparameter tuning approaches and performance with available references are discussed for the data under investigation
- Model extension from lab-scale to pilot-scale successfully demonstrated

Keywords artificial neural network-multilayer perceptron (ANN-MLP) · data quality · machine learning · NIR spectroscopy

Abbreviations

ANN-MLP	Artificial Neural Network-Multilayer Perceptron
API	Active Pharmaceutical Ingredient
AR	Adaboost Regression
BR	Bagging Regression
BU	Blend Uniformity
BVD	Bias-Variance Decomposition

✉ Amrit Paudel
amrit.paudel@tugraz.at

✉ Gobi Ramasamy
gobi.r@christuniversity.in

Extended author information available on the last page of the article

CR	Catboost Regression
CU	Content Uniformity
CV	Cross-Validation
DW-Test	Durbin-Watson test
DQM	Data Quality Metrics
DT	Decision Tree
EDA	Exploratory Data Analysis
EMSC	Extended Multiplicative Scatter Correction
ETR	Extreme Tree Regression
FT	Fourier Transform
GIGO	Garbage In Garbage Out
GBM	Gradient Boosting Machine
HPLC-UV	High-Performance Liquid Chromatography Ultraviolet
iPLS	Interval Partial Least Squares (iPLS)
KNN	K-Nearest Neighbors
LightGBM	Light Gradient Boosting Machine
LOOCV	Leave One Out Cross-Validation
LoR	Logistic Regression
LR	Linear Regression
MAE	Mean Absolute Error
MSC	Multiplicative Scatter Correction
MSE	Mean Square Error
NIR/NIRS	Near-Infrared Spectroscopy
PAT	Process Analytical Technology
PC	Principal Component
PCA	Principal Component Analysis
PLS	Partial Lease Squares
R^2	Coefficient of Determination
CI	Confidence Interval
RF	Random Forest
RMSE	Root Mean Square Error
RTRT	Real Time Release Testing
SA	Sensitivity Analysis
SG	Savitzky-Golay
SNV	Standard Normal Variate
UV-Vis	Ultraviolet or Visible
VIF	Variance Inflation Factor
XGB	Extreme Gradient Boost

Introduction

Since near-infrared spectroscopy (often referred to as NIR or NIRS) is non-destructive and non-intrusive, it requires little to no sample preparation. Moreover, its overall analysis time may be considerably reduced hence it is an ideal real-time analytical tool [1–4]. Few sectors employ NIRS to effectively evaluate both chemical and physical features of solid samples which includes fine chemicals, agriculture, the food and dairy industry, pharmaceuticals, cosmetics, pulp and paper, three-dimensional (3D) printing, and precision medicine [5–14]. Derivatization, normalization, scatter correction, and advanced approaches (sequential and parallel) are a few of the data pre-processing techniques used to conceal physical information and retrieve chemically related information from NIR data [15–18]. Modelling is employed after physical/chemical information has been segmented. If underlying linearity assumptions are satisfied, multivariate linear calibration models will frequently perform better than the non-linear models [14, 19, 20]. Principal component regression (PCR) and partial least squares regression (PLS) is used in multivariate linear models [14, 19]. Since NIR peaks are frequently broad, chemical properties are extracted using resolution enhancement techniques like derivatization, difference spectroscopy, deconvolution, two-dimensional (2D) correlation spectroscopy, self-modelling curve resolution methods, machine learning (ML), and deep learning (DL) [21–23].

For NIR-based BU or CU, current pre-processing approaches involving normalization, second derivative, multivariate scatter correction (MSC), extended MSC (EMSC), and standard normal variate (SNV) might not be pertinent if any of the following conditions exist: (1) data demonstrating heteroscedasticity, non-normality, and multicollinearity which are the prerequisites for the parametric models [20, 24, 25], (2) also, if physical properties are of simultaneous importance, (3) the types of physical fluctuations are uncontrollable or may not be representative of future samples, (4) the artifacts are non-linear/non-parametric. To the best of our knowledge, this is the first work to provide an intuitive understanding of the gaps in existing NIR-based BU analysis.

The study examines data from pharmaceutical tablets manufactured in labs, pilot plants, and at production levels. The data contains 310 tablets NIR spectra that were evaluated over 404 wave numbers between 7400 and 10,500 cm^{-1}. The tablet dataset is freely accessible at http://www.models.life.ku.dk/research/data/Tablets/. A good collection of work has been done using the "Tablet" dataset, including non-linear calibration [19], deep learning approaches for regression and classification [26, 27], and variable importance selection [28, 29] methodologies to support chemometrics, machine learning, and deep learning. However, most of the above-mentioned work compared the models' performance measures. Similarly, most researchers intend to use an empirical method for choosing a subsample while using an open dataset. However, since selecting sub-samples from a dataset is a critical step, we attempt to describe a workflow. A portion of the tablet data was used to create the models in this study, and performance tests using hybrid cross-validation and external cross-validation were implemented. The output of the PLS model is the benchmark used for comparison. Although ANN models generate poor predictive performances, we hypothesize that the generalizability and transferability of ANN models will be better than the linear PLS models. To achieve the aforementioned objective, we employ the following multidisciplinary approach.

a. **Advanced Data Visualization**: Approaches to select the sub-samples for model development
b. **Quality Metrics:** To understand the deviations from the linear assumptions
c. **Model Selection and Hyperparameter Tuning**: For Artificial Neural Network
d. **Performance Metrics:** Hold out, Internal Cross-validation (CV), external/hybrid cross-validation, and Bootstrapping
e. **External Validation:** Develop Model on Lab Batch then Extend to Pilot and Production batches

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 [30]) from the literature was utilized. Tablet [24] 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^{-1}. The tablets were manufactured at the laboratory (lab), pilot plant, and production scale. For more detailed information, please refer to the sub-section literature review under the 'Results' section.

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), where each was used to assess the model's predictive power [31, 32]. 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).

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

$$MAE = \frac{1}{n} \Sigma_{i=1}^{n} \left| y_i^{actual} - y_i^{pred} \right| \quad (2)$$

$$MSE = \frac{1}{n} \Sigma_{i=1}^{n} \left(y_i^{actual} - y_i^{pred} \right)^2 \quad (3)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \Sigma_{i=1}^{n} \left(y_i^{actual} - y_i^{pred} \right)^2} \quad (4)$$

where y_i^{actual} and y_i^{pred} represented the reference and ML predicted values, respectively. On the other hand, $y^{actual,mean}$ 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.

Hold-out Dataset: Assess Selection of Best ML Model

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.

K-fold Cross Validation (CV): Model Generalizability

Sensitivity analysis identifies the causes of bias and variance in model inputs by examining how they differ from the model's output. Point estimates obtained from hold-out data are regarded as being ambiguous, hence sensitivity analysis is considered to be desirable. Resampling techniques like k-fold CV are among the most affordable ways to undertake sensitivity analysis on the generated model. This method divides the given dataset into k groups, each of which can be utilized as a training set while the other groups serve as the test set [33].

External/Hybrid Validation: Model Transferability

The external validation sample could be made up of brand-new pilot or production-scale samples. This is precisely how the model lifecycle management can be established. Additionally, a novel strategy known as internal–external validation architecture, is utilized by Muthudoss *et al.*, [34]. This strategy combines the advantages of internal and external validation. The model performance on production scale batches is predicted using these two approaches.

Data Analysis and Statistics

For this study, we explored the tablets datasets acquired using NIR as described in Ref [30] (found at: 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 the LightGBM package (version '3.2.1') [35], Xgboost package (version '1.5.2') [36], Catboost package (version '1.0.4') [37, 38] and sci-kit sklearn package (version '0.24.0') [39] in Python. Matplotlib package (version '3.4.1') [40] were employed in

generating plots. Statistical analysis and visualization were carried out using JMP standard package (JMP®, Version 16, SAS Institute Inc. Cary, NC, 1989–2022).

Results and Discussion

Original and Related Work

In the original work [30], the authors included many parameters or variabilities like different doses in milligrams, different shapes, and sizes of dosage forms, slots in tablets, and three different manufacturing scales (lab, pilot, full/production), coating differences (coated *vs* uncoated), while manufacturing tablets. The NIR was acquired from tablets which represents a wholesome variability. In the original work, the reference analytical tool employed was high performance liquid chromatography whereas the orthogonal tool employed was Raman spectroscopy. Similarly, the authors have used extensive preprocessing followed by partial least squares (PLS)/interval PLS (iPLS) along with tenfold cross-validation [30]. In this study, we demonstrate the generalizability and transferability of the developed model on an EDA-selected data. Moreover, another aim was to demonstrate how exploratory data analysis coupled with machine learning can provide a low RMSE or MAE model compared to the model developed using only traditional chemometric approaches on NIR data.

Data Visualization

A methodical data visualization strategy was implemented because it was clear from the original work that the authors had employed a very complex system, and because it is widely believed that even a slight difference can affect the quality of the data and how it is interpreted. The goal is to select a subset of data that potentially represents the population, i.e., incorporate the bulk of the variations, and then extend the analysis to new samples. The first part of the paper focuses on identifying the subset (sample) from the overall dataset (population).

The count of tablets is shown in Fig. 1, with the type shown on the x-axis and the scale of manufacturing scale on the y-axis while reversed in Fig. 2. The API content in %w/w is depicted as the secondary x-axis, the colour bar, and the colour gradient of Figs. 1 and 2. Only Type-A has the lowest API content (4–6% w/w), which is manufactured at all scales. On the other hand, lab-scale batches of Type B-D typically have the most significant API content, which ranges between 8 and 10% w/w. The pilot and full-scale consist of only a few tablets of higher API content. Lower and higher API content can alter a tablet's characteristics and can also contain common irregularities like noise or NIR spectral variations. Hence, it is recommended as well as inferred scale-based sample selection be carried out. Next, visualizations are carried out to address whether the subset

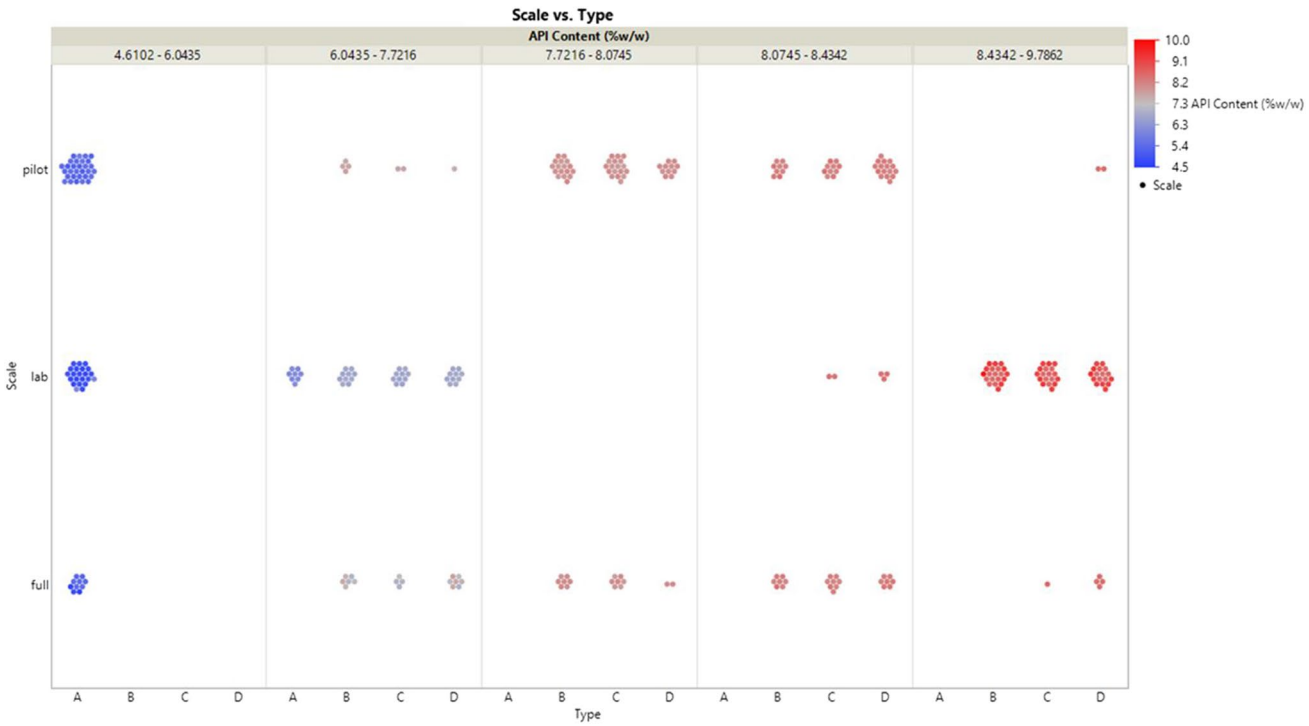

Fig. 1 Approaches to understand the data (scale *vs*. type)

of samples is selected from lab/pilot/full scale batches. Visualizations for the comparison of the scale of manufacturing and the variations were carried out. Figure 3 is a boxplot demonstrating the variations among the scale of samples as a function of type of manufacture and API content (%w/w). Similarly, Fig. 4 represents a vertical bar chart depicting the manufacturing scale as a function of API content (%w/w).

Data Quality Metrics (DQM)

Testing Parametric Assumptions

The quality of input data comprehends the quality of the model, which is mainly referred to as the "GIGO or garbage-in garbage-out" concept [24, 25, 34]. To this end, the tablet dataset was put through several rigorous tests to look for anomalies such as outliers, linearity, multicollinearity, homogeneity in variance, or homoscedasticity, and clustering tendency. The assumptions to be satisfied for the parametric test and their results are shown in Table I. For details about the procedures please refer to [24, 25, 34].

Diagnostic Data Analysis

Reiterating, NIR spectra are sensitive to both physical and chemical information. Such datasets require a significant, visible reduction in dimensionality while minimizing information loss. Various statistical, chemometric, machine learning, and deep learning-based data analytics methodologies have been used to decoded individual contributions [34]. One of the earliest and most prominent methods for this purpose is principal component analysis. It generates new, uncorrelated variables and it is an a *priori* strategy (unsupervised approach) that gradually maximizes the variance of the dataset. Since it preserves as much statistical data (or "variability") as possible, it is an adaptive data analysis technique. The combination of NIR and PAT has been successfully implemented to comprehend the operational processes of the milling operation [41], decipher the effects of powder sampling on undesired segregation of particles [24, 42, 43], and monitor real-time release testing [44, 45].

We employ advanced visualizations and PCA as the exploratory diagnostic tool. In order to determine the randomness, representativeness or segregation error, procedure devised by Saravanan *et al*., [24] was utilized. Figures 5a and b show the PCA results for this, which were then used to extrapolate more details about the Tablet dataset variability and likely selection of subsamples. No peaks appeared or disappeared when the tablet dataset (310 NIR spectral observations) was superimposed (data not shown). Without any prior data pretreatment, these tablet datasets were exposed to PC projection and score plot interpretation. This process increases dataset variance, which is useful for capturing variability associated with physical properties. The variability in the dataset is adequately captured by using just two principal components as demonstrated using a scatterplot. The PCA-based

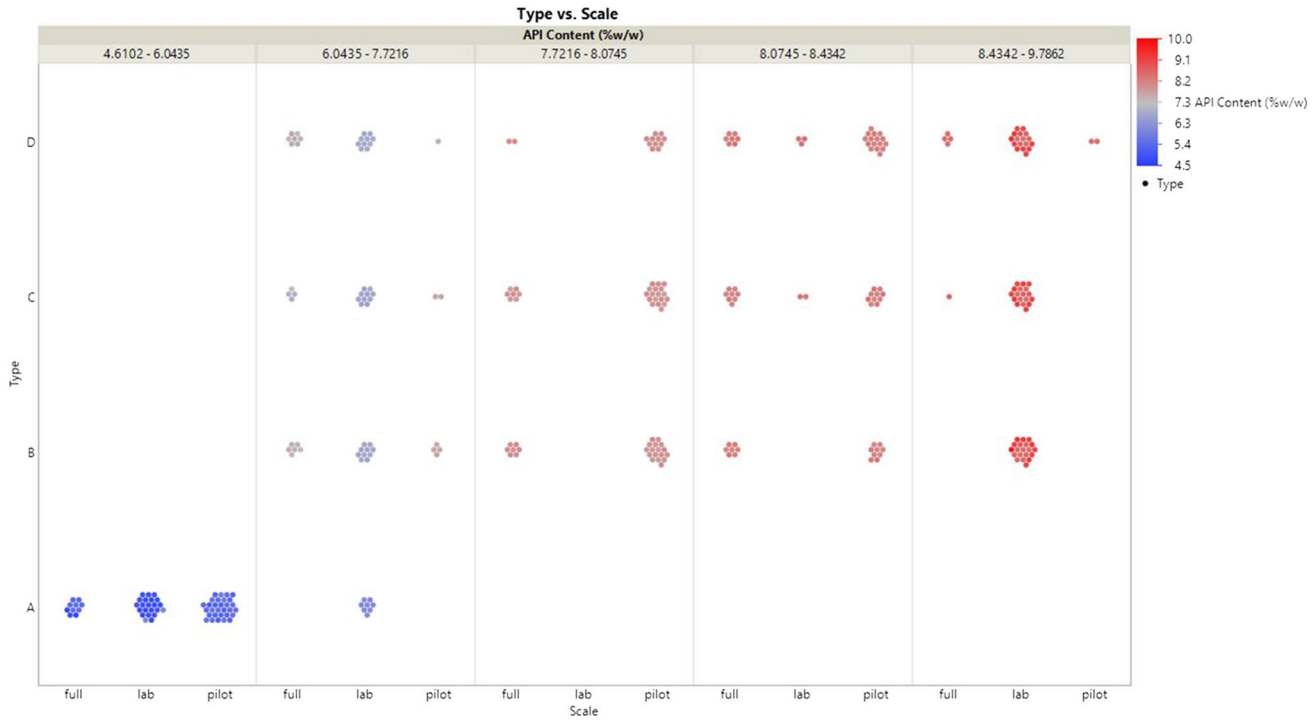

Fig. 2 Approaches to understand the data (type *vs*. scale)

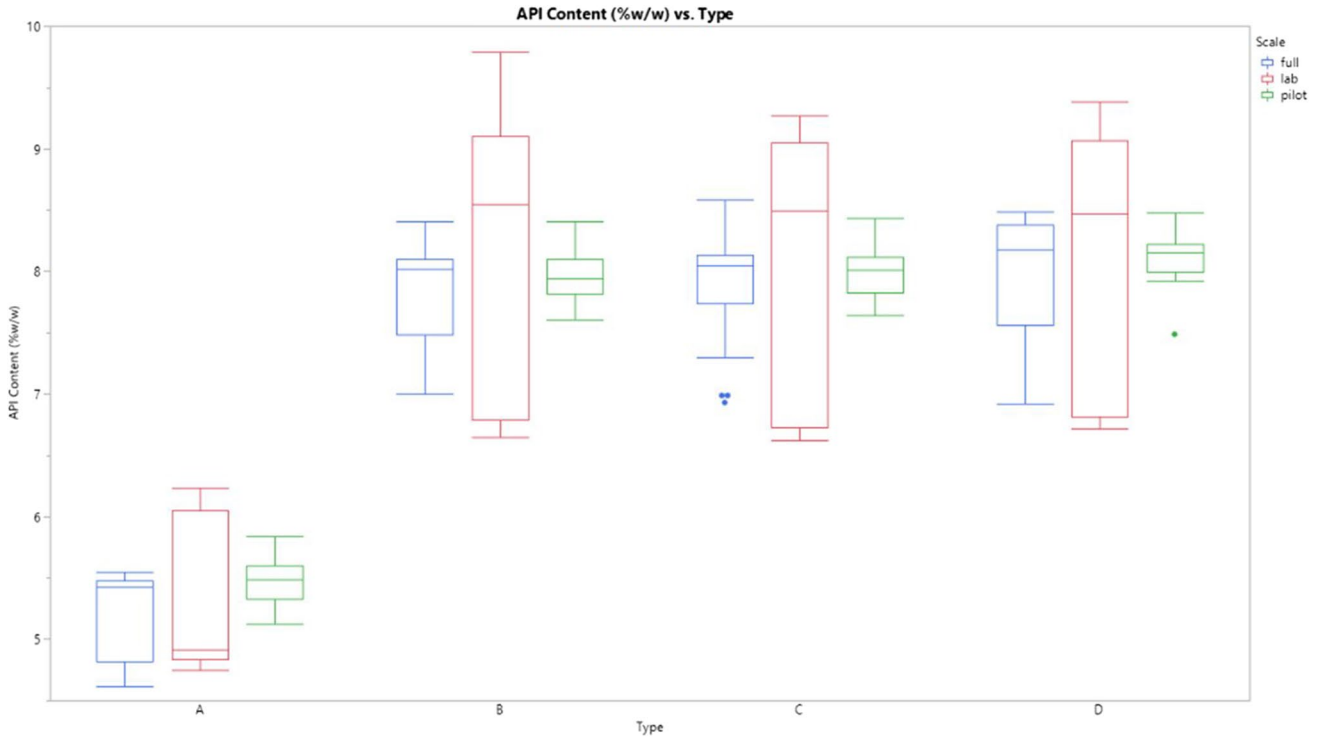

Fig. 3 API content (%w/w) as a function of type (A, B, C, D)

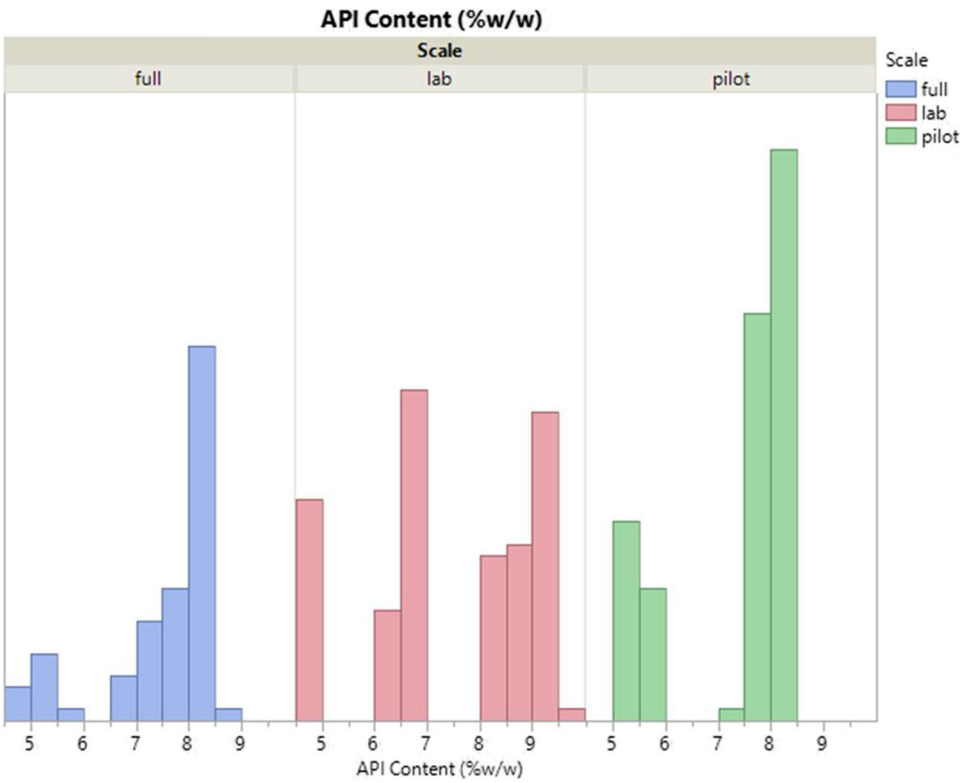

Fig. 4 API content (%w/w) as a function of full scale, pilot scale and lab scale

Table I Data Quality Metrics

Test	Decision rule	Results
Shapiro–Wilk Test (test for normality)	Null hypothesis: $p > 0.05$, data follow normal distribution	p-value < 0.05, Non-normal
	Alternate Hypothesis: p-value < 0.05 implies data do not follow normality	
Levene's Test (test for homogeneity of variance/homoscedasticity)	Null hypothesis: $p > 0.05$, data shows homoscedasticity	p-value < 0.05, Heteroscedasticity exist
	Alternate Hypothesis: p-value < 0.05, data is not show homoscedastic	
Scatterplot Matrix (SPLOM) (test for linearity)	Pairwise combinations of continuous variables	Figure S1. Independent features are non-linear with target cell
Variance Inflation Factor (VIF) (test for multicollinearity)	Values < 5 indicate less or no,	Figure S2. Values > 10, Multicollinearity exist
	Values 5–10 imply moderate	
	Values > 10 indicate severe	
Cook's Distance	A datapoint that has extreme values (either large or small) than nearest value	Figure S3. Outliers and highly influential data points observed
Durbin-Watson (measure for autocorrelation and test for independency)	Null hypothesis: $p > 0.05$, implies data has no first order autocorrelation	Figure S4 and S5. Figure p-value < 0.05, Serial Autocorrelation exist, and Data are not independent. Data is statistically nonlinear
	Alternate Hypothesis: p-value < 0.05, implies first order autocorrelation exists	

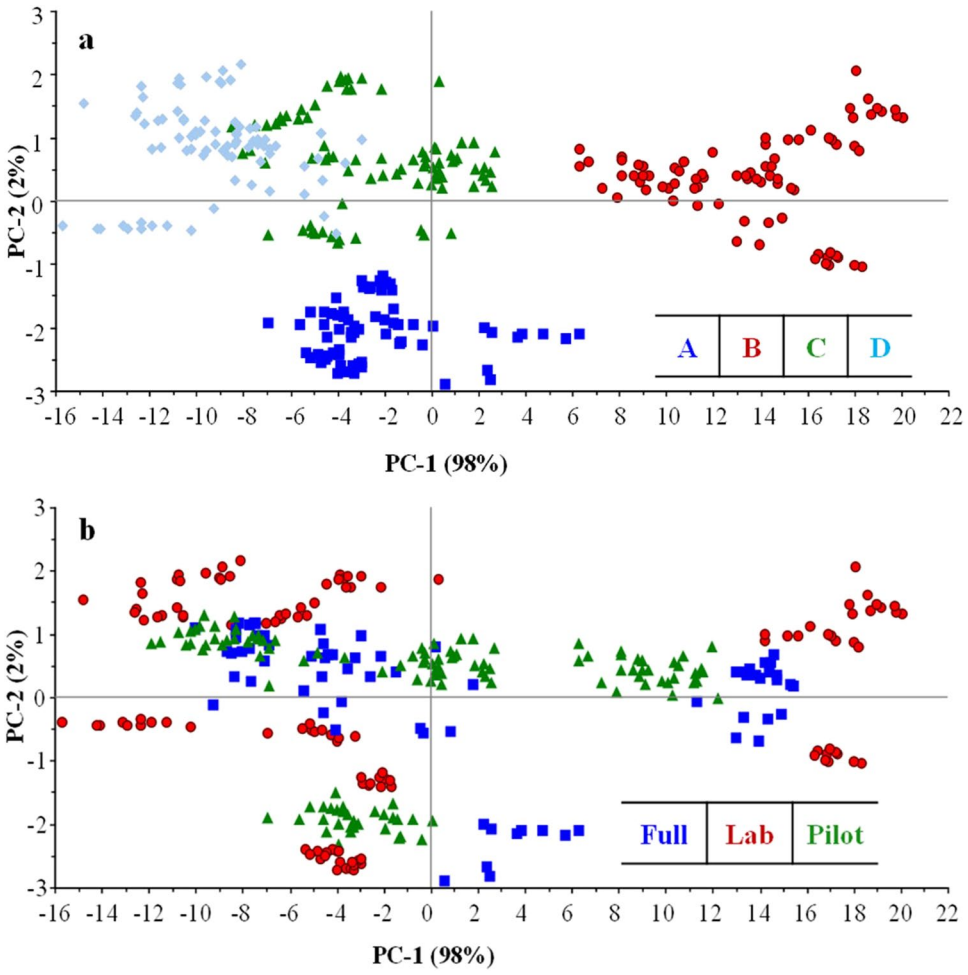

Fig. 5 **a** Exploratory data analysis employing PCA to determine pattern (scores plot based on tablet dataset with respect to type). **b** Exploratory data analysis employing PCA to determine pattern (scores plot based on tablet dataset with respect to batch)

score plots for type (Fig. 5a) and batch (Fig. 5b) were used to analyze tablet datasets in order to diagnose patterns and trends. Data points in the scatterplot were found to be well-clustered and distinct trends could be detected, indicating that there are differences between and within samples. The next stage is to choose the subsamples for developing predictive machine learning models based on these interpretations and in conjunction with advanced visualization. The best dataset that could mimic population could be lab scale data with all types (approach 2) rather than approaching vice-versa. These various visualizations and diagnostic PCA plots make it abundantly evident that the lab batches have the widest variability and can serve as the ideal samples for predictive modelling, which can also account for potential future variations.

Inferences from DQM

Evaluating for the presence of extreme values, inhomogeneity of variance, correlated independent variables, non-independence of data is considered a prerequisite as they can contribute to the non-linearity associated with unprocessed data [24, 34]. Hence, the data quality metrics (DQM) involving test for normality, test for heteroscedasticity, test for multicollinearity, test for outliers were carried out, refer to Table I. The results indicated that these assumptions were violated, which can yield inaccurate parameter estimates. That is, it is regarded as a significant concern when a model is trained on one dataset (in this case, lab-scale) and forecasted to another with a distinctive or unknown structure of collinearity, heteroscedasticity, non-independence, noise, or outlier (like pilot or full/production size). Additionally, the Durbin-Watson (DW) test [46–48] was utilized to statistically evaluate for the presence of non-linearity, assess autocorrelation, and test for data independence, Table I. Since the p-value < 0.05, presence of serial autocorrelation is confirmed, and it can be concluded that the data are not independent. Through the DQM analysis, it is evidenced that the model open-source data was found not to comply with the linearity assumptions and the data is non-linear. Hence, apart from PLS we selected a range of non-linear regression tools from distance-based approach (kNN), decision-tree (bagging and boosting), support vector machine etc. However, these models also did not perform well. Hence, artificial neural network-multilayer perceptron incorporating a hybrid approach involving non-linear architecture followed by linear hyperparameters for API content (%w/w) estimation was employed.

Selection of Subset of Sample During Analytical or Product Development Lifecycle

The main concept of the data visualization as well as DQM is to choose the suitable subset of samples for model training, validation and subsequent analysis. This is significant since the method development is based on lab batches at the beginning of product development in the R&D phase. The methodology created should be a generalised approach with the expectation that it will be scaled to forecast the pilot and production scales or subsequent batches of samples. It is not advisable to keep tuning the parameters of the model as the formulation changes and/or samples changes occurs. That is, the developed model needs to be robust to detect the actual variabilities than the random variations or noises. To understand as well as extract the changes existing between the batches, we first adopted the advanced visualization strategy, then the model development was carried out using the lab batches as training set. The pilot and production scales are utilized as external validation samples to demonstrate the model generalizability and model transferability. Target variable (API %w/w) was visualized as a function of scale as an independent variable. The various data visualization strategies, like boxplot, histogram, scatterplot, scatterplot matrix, DQM, diagnostic plots etc., were performed. The results indicated that the lab scale batches consisted of the maximum variabilities as shown in Fig. 3. Moreover, the data is found to form clusters or groups as shown in Fig. 5a and b.

Model Selection

Univariate Analysis

This data was then subjected to univariate analysis using the API specific peak 8833 cm^{-1} as shown in Fig. 6. It is well known that the NIR results are prone to physical variations specifically scattering variations hence the nonperformance can be attributed to these causes.

Model Selection: Hold-out Dataset

- ***Traditional Machine Learning***

Reiterating, DQM from Table I as well as Figs. 7, and 8 demonstrate the non-linearities associated with the "Tablet" dataset. Hence, as a first step, a framework for choosing an appropriate baseline model is employed, which includes both linear and non-linear models. The framework selection is performed on the hold-out dataset (for details refer to section on "Training-Test Split"). The calibration data, also known as "hold-out data", are divided in this study using a random selection approach in the ratios of 80% (for the train): 20% (for the test). By separating training data from test data, which precisely reflect the calibration data, unbiased evaluation of the machine learning algorithms performance is made possible. The top-performing models were ranked using the performance assessment measures R^2,

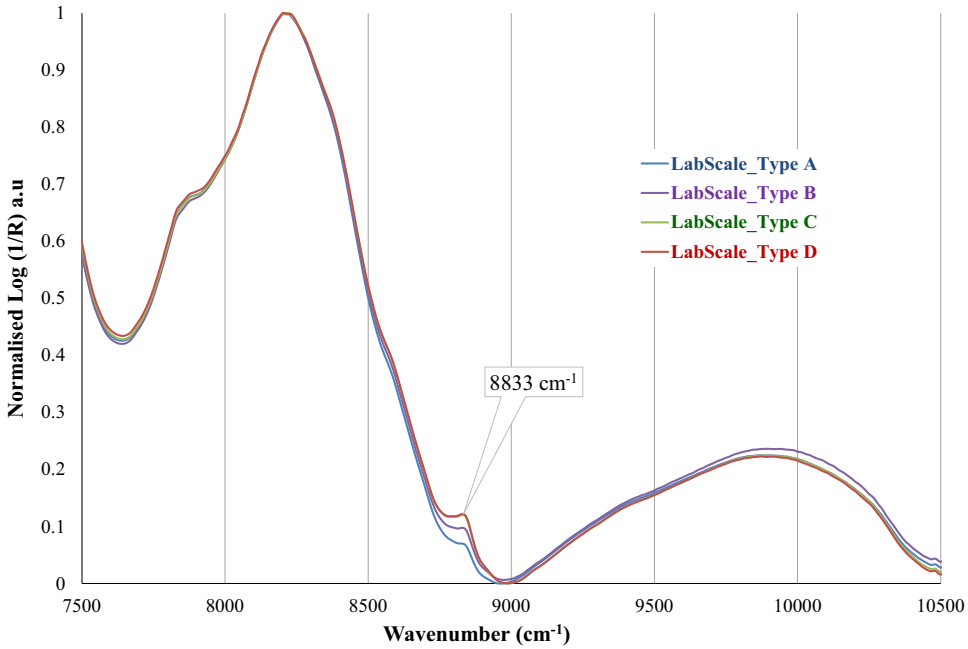

Fig. 6 Typical NIR spectra of lab scale samples

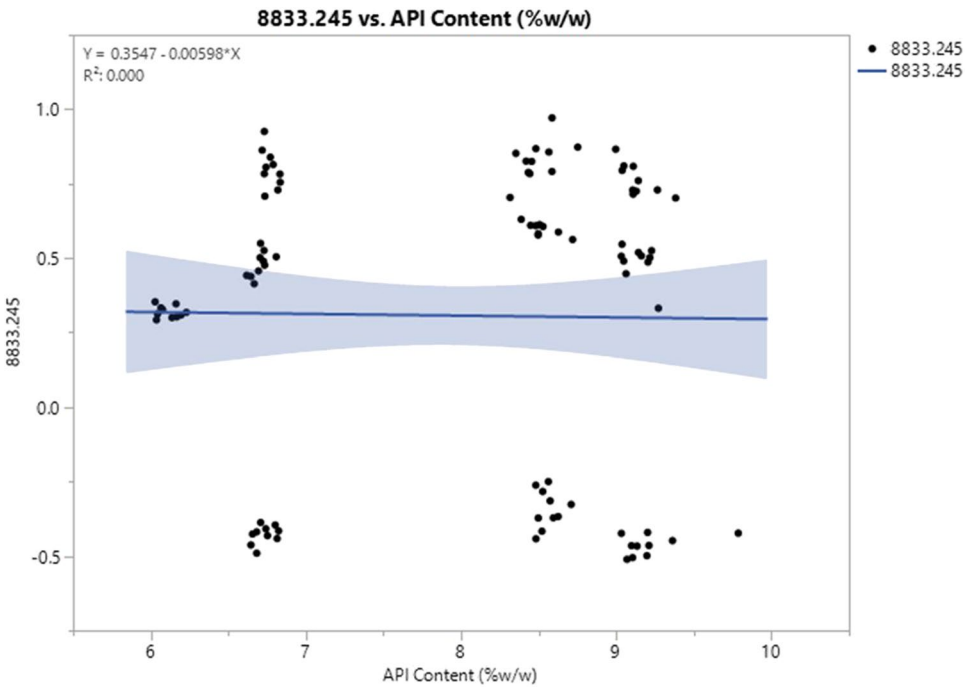

Fig. 7 Univariate peak analysis based on API content *vs*. API peak at 8833 cm^{-1}

MAE, MSE, and RMSE. Table II summarizes the results of numerous models (the top 10) and performance measures. As reported by various authors as well as it is evident from the table that the linear models like PLS are comparatively efficient, refer to Table II.

PLS is the de facto linear multivariate regression approach with a wide range of analytical applications in the pharmaceutical industry. It is well known that the approach is capable to handle weak nonlinearities in the independent variable (X) dimension, it lacks to link this relationship between independent (X) and dependent (response or y) variables [19]. Similarly, although PLS is the de facto standard for linear multivariate regression, the pre-processing of the spectral data is considered a prerequisite. Savitzky-Golay (SG) based derivatization (first or second derivative), normalization (min–max or vector or similar normalization), standard normal variate (SNV), multiplicative scatter correction (MSC), and extended MSC (EMSC) are commonly used pre-processing approaches [15–18]. However, these approaches are parametric in nature, that is, these work well

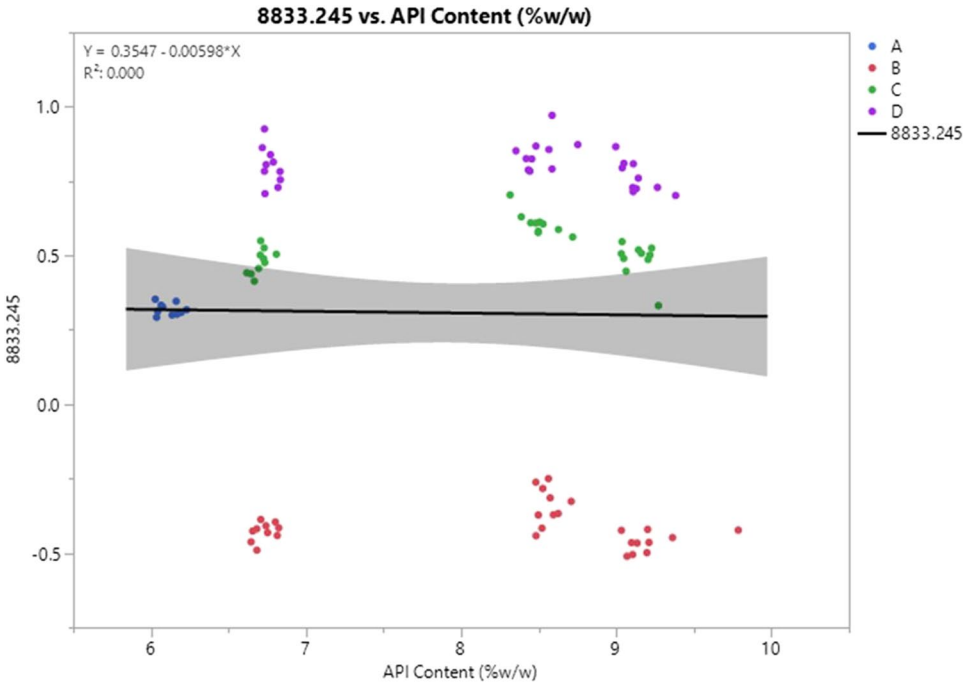

Fig. 8 Univariate peak analysis based on API content *vs*. API peak at 8833 cm^{-1} (addition of another dimension denoting type depicts both sub-clustering and non-linearity)

Table II Comparison of the Various Model as a Function of Performance Metrics

	Model	R^2_Train_Score	R^2_Test_Score	MAE_Train_Score	MAE_Test_Score	MSE_Train_Score	MSE_Test_Score
3	PLS Regression	0.982199	0.967942	0.158755	0.208416	0.041466	0.090002
6	Decision Regression	1.0	0.879293	0.0	0.250635	0.0	0.338878
8	Bagging Regression	0.946867	0.925163	0.228638	0.299584	0.123771	0.2101
11	XGB Regression	0.999999	0.906188	0.000778	0.304566	0.000001	0.263372
1	Ridge Regression	0.925343	0.937921	0.320858	0.310998	0.173909	0.174284
14	ETR Regression	1.0	0.89311	0.0	0.312222	0.0	0.300086
10	GradientBoost Regression	0.997975	0.881278	0.051976	0.340058	0.004717	0.333305
13	CatBoost Regression	0.99671	0.892106	0.06295	0.350341	0.007664	0.302904
7	Random_Forest Regression	0.962479	0.893493	0.206278	0.395863	0.087404	0.29901
5	KNN Regression	0.779722	0.798693	0.5072	0.540402	0.513128	0.565156
12	LightGBM Regression	0.77531	0.748262	0.551061	0.650319	0.523405	0.706739
9	ADABoost Regression	0.857663	0.78076	0.516457	0.667102	0.331568	0.615503
0	Linear Regression	1.0	0.447688	0.0	0.80399	0.0	1.55058
4	SVM Regression	0.243182	0.252879	1.03347	1.255004	1.762975	2.097493
2	Lasso Regression	0.0	-0.010742	1.387464	1.544971	2.329458	2.837591

for linear and normal data. For NIR-based measurements, current pre-processing approaches involving normalization, second derivative, MSC, EMSC, and SNV might not be pertinent if any of the following conditions exist: (1) data demonstrating homoscedasticity, normality, absence multicollinearity, absence of outliers, absence of non-linearity are the prerequisites for the parametric models [20, 24, 25], (2) also, if physical properties are of simultaneous importance, (3) the types of physical fluctuations are uncontrollable, (4) may not be representative of future samples, (5) the artefacts/residuals/data are non-linear. The presence of noise and outliers cannot be reproducible, hence second derivative and normalization could have erroneous results. That is, the PLS can perform well with internal validation data. However, when presented with new unseen data or new data from a different process etc. might have severe limitations. Hence, a non-linear model like ANN was also evaluated. Additionally, to evaluate tablet dataset, artificial neural network-multilayer

perceptron (ANN-MLP) was employed. Furthermore, comparison between PLS and ANN-MLP with respect to model generalisability and model transferability is evaluated.

- ***Artificial Neural Network-Multilayer Perceptron (ANN-MLP)***

It is hypothesized that there could exist a nonlinear relationship between the sample and the NIR spectra [46]. Because the open-source tablet dataset violated the DQM tests and the existence of non-linearity is established, it has been deemed appropriate to explore beyond the PLS, for more information, refer to Table I. In such scenarios, proceeding with parametric methods could be detrimental. That is, violation of these tests will make the model less reliable, and the predictive results will have more uncertainty and errors. Such mishaps could impact models' explainability, interpretability, generalizability, and transferability. This is one of the discoveries in our previous works [24, 25, 34]. Also, when there are more dependent variables to be investigated, ANN-MLP could be quite useful [46]. Moreover, the currently available spectral pre-processing approaches like SNV, MSC, EMSC, normalization etc. are parametric in nature. One of the objectives was to avoid the preprocessing as that can negatively affect the performance of developed PLS models. Similarly, one other objective was to demonstrate the model generalizability and transferability from lab-scale to pilot-scale to full/production-scale, it is deemed necessity that developed model requires minimum assumptions as well as stable to non-linearities, outliers etc. This alternative is what is discussed in this paper through the application of hybrid linear-nonlinear ANN-MLP framework.

Handling nonlinearity in data using ANN-MLP can be achieved through the careful selection of appropriate architecture (hidden layer structure) and hyperparameters. ANN algorithm can approximate any linear or nonlinear relationship. In general, ANN consist of an input layer, one or more hidden intermediate layers, and output layers [27, 46, 47]. The hyperparameters employed in this study were as follows: (i) solver employed was 'lbfgs' a weight optimizer in the family of quasi-Newton methods (non-linear hyperparameter), (ii) learning rate: 'invscaling' to schedule weight updates, (iii) activation function (linear hyperparameter): 'identity', no-op activation, useful to implement linear bottleneck, (iv) max_iter:125, (v) hidden_layer_sizes = (4,125,8). These were chosen through a few trials and an error-based approach. However, in the future, hyperparameter tuning will be employed. Various parameters like tablet shape, slot, active or API content, type, batch etc. severely impacts the data analysis. The original paper's authors used various pre-processing approaches to overcome these nonlinearities. However, recently, Muthudoss *et al.*, [34] had indicated that even when the changes concerning material, method, analyst, and environment are kept constant, the adoption of data analysis preprocessing and processing procedures could contribute to non-linearities. In this context, it is not sure if the PLS-based approach can solve non-linearity. Hence, a nonlinear model based on multilayer perceptron is implemented. The performance of the nonlinear ANN-MLP on hold-out dataset are satisfactory.

Model Robustness: Generalizability

To evaluate the robustness/generalization capability, sensitivity analysis like k-fold cross-validation (tenfold, bootstrapping and leave one-out) were performed [48–50]. As with hold-out validation, tenfold cross-validation of MLP had lower error rates (mean of 0.2287% w/w ± standard deviation of 0.074% w/w). On the other hand, tenfold CV error rates for PLS were mean of 0.2208% w/w ± standard deviation of 0.061% w/w. Emphasizing that near-infrared spectra are known to be sensitive to physical artefacts, and the detected variations can be attributable to the potent scattering effects of powders [51]. In summary, the model generalization error of CU ± 0.074% w/w was tolerable and suggested that the ANN-MLP model is robust [51].

Model Lifecycle Management and Maintenance: Transferability (External Cross-Validation)

It is widely known that estimates are susceptible to variability or heterogeneity in the dataset, hence inconsistencies. Hence, there determining the model's transferability, recalibration, and/or maintenance needs to be carried out. In this study, two new samples that are quite different from lab scale batches were included. These samples were pilot-scale and full/production scale batches. These samples were termed external validation samples and were subjected to both ANN-MLP and PLS pretrained models. The prediction error results are depicted in Fig. 9a and b. The errors on new samples were expected to be between 0.1548 and 0.3028% w/w for MLP whereas between 0.161 and 0.281% w/w is expected for the PLS baseline model. The ANN-MLP performs in line with the cross-validation results with respect to the pilot-scale batch. However, its performance is slightly less better on full/production scale. This could indicate the impact of unseen variability existing in the full/production scale batches. The future work will be directed to developing a slightly different cross-validation architecture. In summary, the ANN-MLP was found to be at par with the PLS baseline model. These two samples, which could resemble data from different vendors or new batches of raw materials/processed samples, were the most accurate approximations. The findings are depicted in Fig. 10.

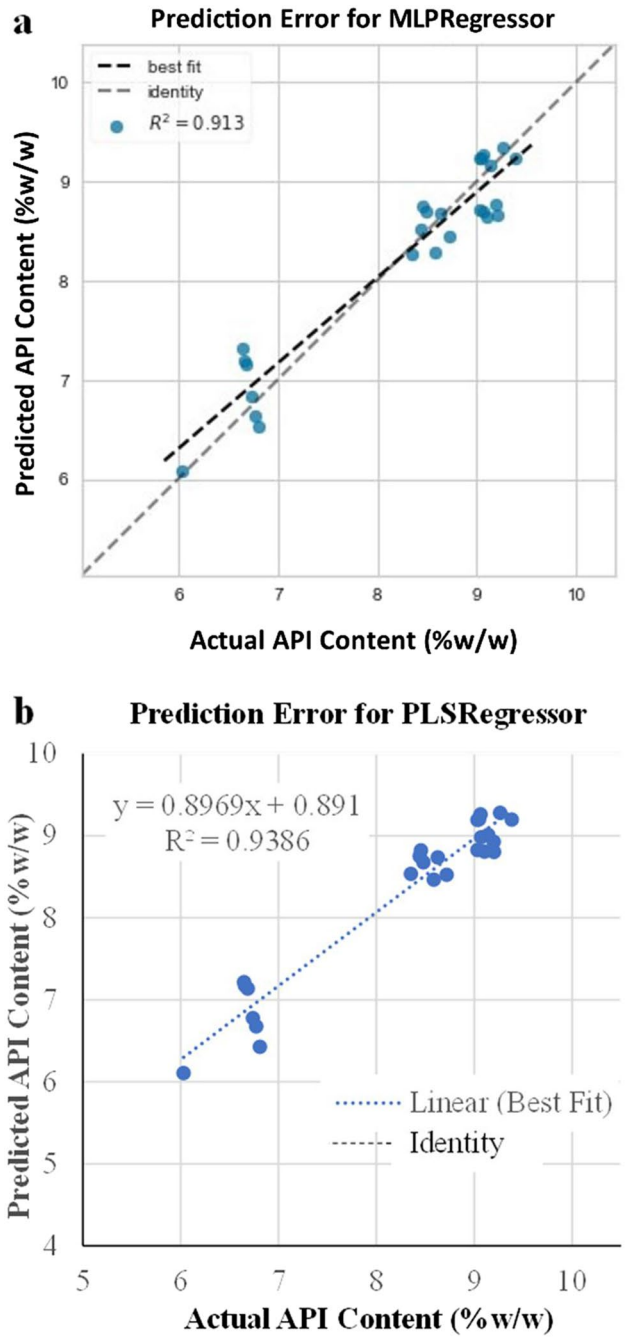

Fig. 9 **a** Prediction error for MLP regressor. **b** Prediction error for PLS regressor

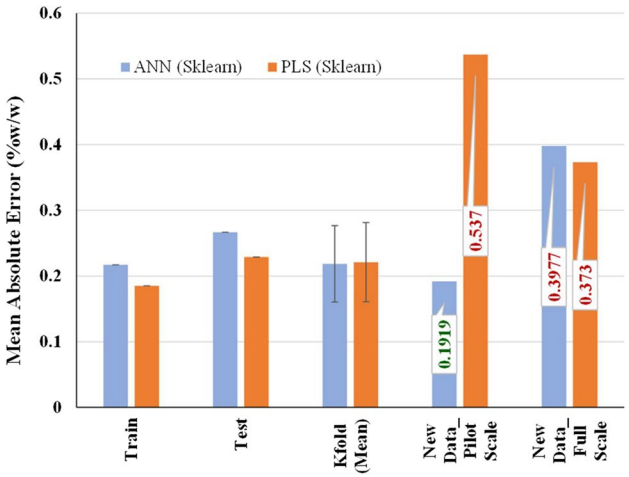

Fig. 10 Comparison of the MAE for train, test, k-fold cross-validation and external validation samples from pilot and full/production scale

This study assessed an open-source "Tablet dataset" that many researchers have thoroughly studied. Most of the available research work focused on developing algorithms, feature engineering, feature selection, etc. This work involved methodical approaches (i) advanced data visualization facilitated the ability to understand different investigated parameters in original work in an explainable way, (ii) augmentation of diagnostic tools made it possible to choose sub-samples with the greatest variability for model development, (iii) inspection of the dataset using quality metrics reveals that it violates linearity assumptions, (iv) based on descriptive, diagnostic analysis and quality metric results, a selection of relevant hyperparameters and their tuning was conducted, (v) model's robustness and lifecycle management were effectively carried out by applying internal hold-out validation, internal cross-validation, and external cross-validation. The model was first developed on lab-batch to evaluate the performance before being expanded to pilot-scale and full-scale (both generalisability and transferability were assessed). In summary, a workflow has been established for the purposes of understanding, curating, choosing sub-samples, developing reliable predictive models, etc. for open-source or publicly accessible data, as illustrated in Fig. 11.

Relative Prediction Error (RPE)

Instead of creating a model by categorising the data according to API content, ANN-MLP models were created on a lab scale and evaluated using complete pilot and full/production scale datasets. The same error, for instance, 0.1508% w/w (lower MAE) and 0.3208% w/w (higher MAE), could have different practical implications for different tablet strengths because errors are direct indicators of a model's performance in this scenario. We therefore give the $y_{nominal}$ error normalised as a function of the tablet's API content (%w/w) and total tablet weight, which is known as the relative prediction error (RPE) in the original study [30]. Table III presents the findings. For each dosage strength, the prediction error ranges between 1.5 and 6.5% w/w at a 95% confidence level. Future work will involve fine-tuning and establishing optimum hyperparameters in order to significantly lower the mean absolute error.

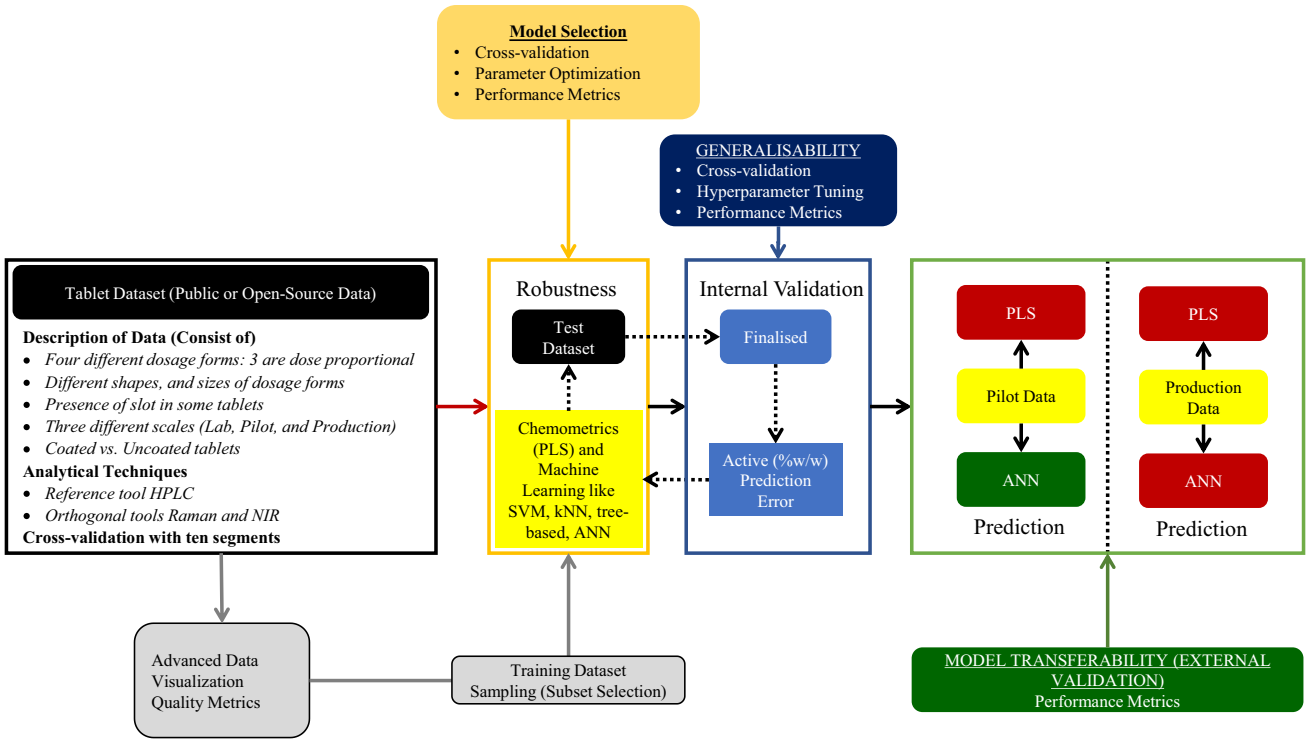

Fig. 11 Workflow for Machine Learning–enabled evaluation of publicly accessible or open-source datasets (Example NIR Spectra of Tablet Dataset)

Conclusion

A methodical process that involves data visualisation and choosing sub-samples from the historical data is described. We describe the performance of ANN-MLP (based on careful selection of hyperparameters) on par with the original PLS work for the first time. The described procedure should be seen as a catalyst for the rapid and efficient implementation of the NIR approach whether it is used offline, online, at line, or inline. In this study, baseline PLS and multilayer perceptron models were employed for regressions analysis on Tablet datasets that were publicly available. The initial work involved in understanding the data from the perspectives of both quality and sample selection front. It was observed that the dataset hosts a lot of

Table III Relative Prediction Error (RPE) for ANN-MLP Model as a Function of Scale

Nominal weight (%)	Nominal content of active substance per tablet (mg)	Nominal tablet weight (mg)	ANN-MLP (Min_MAE) 0.1508 (%w/w)	ANN-MLP (Max_MAE) 0.3028 (%w/w)
Laboratory scale				
4.8	4.3	90	3.1%	6.3%
6.3	5.7	90	2.4%	4.8%
6.9	9.3	125	2.2%	4.4%
9.1	11.4	125	1.7%	3.3%
6.9	12.9	188	2.2%	4.4%
9.1	17.1	188	1.7%	3.3%
6.9	17.3	250	2.2%	4.4%
9.1	22.8	250	1.7%	3.3%
Pilot scale and full scale				
5.6	5	90	2.7%	5.4%
8	10	125	1.9%	3.8%
8	15	188	1.9%	3.8%
8	20	250	1.9%	3.8%

variabilities hence advanced visualisation helped in selecting the right sample that can be representative of the population. Based on this understanding, lab batch was selected and the PLS and MLP models were trained. Both the models were found to be generalizable based on k-fold. The pretrained models were then employed on new data from pilot and full/production scale to understand the model life cycle and transferability. MLP model demonstrated superior transferability concerning pilot scale while it was at par with full/production scale. Understanding and overcoming such differences will be the subject of future paper. A method for meticulously choosing sub-samples from historical datasets is developed using cutting-edge visualization, hyperparameter tuning, and cross-validation. This complex methodology aids in bringing non-linear models' performance up to par with that of linear PLS models.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1208/s12249-022-02493-5.

Acknowledgements HA, PM acknowledge on 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/).

Author Contribution **Hussain Ali:** Conceptualization; Methodology; Formal Analysis; Writing, Original Draft; Writing, Review and Editing.

Prakash Muthudoss: Software; Validation, Methodology; Validation; Formal Analysis: Writing, Review and Editing.

Manikandan Ramalingam: Resources, Review and Editing.

Lakshmi K.: Resources, Supervision.

Amrit Paudel: Supervision, Writing—Review and Editing.

Gobi Ramasamy: Supervision, Project Administration, Writing—Review and Editing, Funding Acquisition.

Funding Open access funding provided by Graz University of Technology.

Data Availability Data is available from authors on request.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. Blanco M, Coello J, Iturriaga H, Maspoch S, De La Pezuela C. Near-infrared spectroscopy in the pharmaceutical industry. Critical review. Analyst. Royal Society of Chemistry; 1998;123:135R--150R.
2. Luypaert J, Massart DL, Vander HY. Near-infrared spectroscopy applications in pharmaceutical analysis. Talanta Elsevier. 2007;72:865–83.
3. Pasquini C. Near infrared spectroscopy: a mature analytical technique with new perspectives–a review. Anal Chim Acta Elsevier. 2018;1026:8–36.
4. Razuc M, Grafia A, Gallo L, Ramírez-Rigo MV, Romañach RJ. Near-infrared spectroscopic applications in pharmaceutical particle technology. Drug Dev Ind Pharm. Taylor \& Francis; 2019;45:1565–89.
5. Okubo N, Kurata Y. Nondestructive classification analysis of green coffee beans by using near-infrared spectroscopy. Foods. Multidisciplinary Digital Publishing Institute; 2019;8:82.
6. 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. Elsevier; 2021;225:121971.
7. 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.
8. Zhu L, Lu SH, Zhang YH, Zhai HL, Yin B, Mi JY. An effective and rapid approach to predict molecular composition of naphtha based on raw NIR spectra. Vib Spectrosc. Elsevier; 2020;109:103071.
9. Liu Y, Fearn T, Strlič M. Quantitative NIR spectroscopy for determination of degree of polymerisation of historical paper. Chemom Intell Lab Syst. Elsevier; 2021;214:104337.
10. Trenfield SJ, Tan HX, Goyanes A, Wilsdon D, Rowland M, Gaisford S, *et al.*. Non-destructive dose verification of two drugs within 3D printed polyprintlets. Int J Pharm. Elsevier; 2020;577:119066.
11. Beć KB, Grabska J, Badzoka J, Huck CW. Spectra-structure correlations in NIR region of polymers from quantum chemical calculations. The cases of aromatic ring, C= O, C≡ N and C-Cl functionalities. Spectrochim Acta Part A Mol Biomol Spectrosc. Elsevier; 2021;262:120085.
12. Cayuela-Sánchez, José A., Javier Palarea-Albaladejo, Juan Francisco García-Martín and M del CP-C. Olive oil nutritional labeling by using Vis/NIR spectroscopy and compositional statistical methods. Innov Food Sci \& Emerg Technol. Elsevier; 2019;51:139–47.
13. Sulub Y, Wabuyele 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.
14. 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. Elsevier; 2021;192:113684.
15. Xiao-Li L, Hua L. Quantitative analysis of amlodipine besylate powder using near infrared spectroscopy combined with partial least-squares. ICAE 2011 Proc 2011 Int Conf New Technol Agric Eng. 2011;874–7.
16. Jiao Y, Li Z, Chen X, Fei S. Preprocessing methods for near-infrared spectrum calibration. J Chemom. Wiley Online Library; 2020;34:e3306.
17. Stordrange L, Libnau FO, Malthe-Sørenssen D, Kvalheim OM. Feasibility study of NIR for surveillance of a pharmaceutical process, including a study of different preprocessing techniques. J Chemom A J Chemom Soc. Wiley Online Library; 2002;16:529–41.
18. Sulub Y, Konigsberger M, Cheney J. Blend uniformity end-point determination using near-infrared spectroscopy and multivariate calibration. J Pharm Biomed Anal Elsevier. 2011;55:429–34.
19. 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. Available from: https://doi.org/10.1016/j.aca.2013.12.002.

20. Mishra P, Rutledge DN, Roger J-M, Wali K, Khan HA. Chemometric pre-processing can negatively affect the performance of near-infrared spectroscopy models for fruit quality prediction. Talanta. Elsevier; 2021;229:122303.
21. Ozaki Y, Šašić S, Jiang JH. How can we unravel complicated near infrared spectra?—Recent progress in spectral analysis methods for resolution enhancement and band assignments in the near infrared region. J Near Infrared Spectrosc. SAGE Publications Sage UK: London, England; 2001;9:63–95.
22. Sadat A, Joye IJ. Peak fitting applied to fourier transform infrared and raman spectroscopic analysis of proteins. Appl Sci. MDPI; 2020;10:5918.
23. Roggo Y, Jelsch M, Heger P, Ensslin S, Krumme M. Deep learning for continuous manufacturing of pharmaceutical solid dosage form. Eur J Pharm Biopharm Elsevier. 2020;153:95–105.
24. Saravanan D, Muthudoss P, Khullar P, Rose VA. Quantitative microscopy: particle size/shape characterization, addressing common errors using 'analytics continuum' approach. J Pharm Sci. 2021;110:833–49.
25. Muthudoss P, Kumar S, Ann EYC, Young KJ, Chi RLR, Allada R, *et al.*. Topologically directed confocal Raman imaging (TD-CRI): advanced Raman imaging towards compositional and micromeritic profiling of a commercial tablet components. J Pharm Biomed Anal. Elsevier; 2022;114581.
26. Jernelv IL, Hjelme DR, Matsuura Y, Aksnes A. Convolutional neural networks for classification and regression analysis of one-dimensional spectral data. 2020; Available from: http://arxiv.org/abs/2005.07530.
27. Acquarelli J, van Laarhoven T, Gerretzen J, Tran TN, Buydens LMC, Marchiori E. Convolutional neural networks for vibrational spectroscopic data analysis. Anal Chim Acta [Internet]. Elsevier Ltd; 2017;954:22–31. Available from: https://doi.org/10.1016/j.aca.2016.12.010.
28. Farrokhnia M, Karimi S. Variable selection in multivariate calibration based on clustering of variable concept. Anal Chim Acta [Internet]. Elsevier B.V.; 2016;902:70–81. Available from: https://doi.org/10.1016/j.aca.2015.11.002.
29. Tran TN, Afanador NL, Buydens LMC, Blanchet L. Interpretation of variable importance in Partial Least Squares with Significance Multivariate Correlation (sMC). Chemom Intell Lab Syst [Internet]. Elsevier B.V.; 2014;138:153–60. Available from: https://doi.org/10.1016/j.chemolab.2014.08.005.
30. 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.
31. Andersen CM, Bro R. Variable selection in regression—a tutorial. J Chemom Wiley Online Library. 2010;24:728–37.
32. Rajalahti T, Kvalheim OM. Multivariate data analysis in pharmaceutics: a tutorial review. Int J Pharm Elsevier. 2011;417:280–90.
33. Yang Y, Ye Z, Su Y, Zhao Q, Li X, Ouyang D. Deep learning for in vitro prediction of pharmaceutical formulations. Acta Pharm Sin B Elsevier. 2019;9:177–85.
34. Prakash Muthudoss, Ishan Tewari, Rayce Lim Rui Chi, Kwok Jia Young, Eddy Yii Chung Ann, Doreen Ng Sean Hui, Ooi Yee Khai, Ravikiran Allada, Manohar Rao, Saurabh Shahane, Samir Das, Irfan Babla, Sandeep Mhetre AP. Machine learning-enabled NIR spectroscopy in assessing powder blend uniformity: clear-up disparities and biases induced by physical artefacts. AAPS PharmSciTech [Internet]. 2022;23. Available from: https://doi.org/10.1208/s12249-022-02403-9.
35. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al.*. Lightgbm: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30.
36. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. Proc 22nd acm sigkdd Int Conf Knowl Discov data Min. 2016. p. 785–94.
37. Dorogush AV, Ershov V, Gulin A. CatBoost: gradient boosting with categorical features support. arXiv Prepr arXiv181011363. 2018.
38. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst. 2018;31.
39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.*. Scikit-learn: machine learning in Python. J Mach Learn Res JMLR org. 2011;12:2825–30.
40. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci \& Eng. IEEE Computer Society; 2007;9:90–5.
41. Gottlieb DM, Schultz J, Bruun SW, Jacobsen S, Søndergaard I. Multivariate approaches in plant science. Phytochemistry Elsevier. 2004;65:1531–48.
42. Alcalà M, Blanco M, Bautista M, González JM. On-line monitoring of a granulation process by NIR spectroscopy. J Pharm Sci Wiley Online Library. 2010;99:336–45.
43. Chavan RB, Bhargavi N, Lodagekar A, Shastri NR. Near infra red spectroscopy: a tool for solid state characterization. Drug Discov Today Elsevier. 2017;22:1835–43.
44. Galata DL, Könyves Z, Nagy B, Novák M, Mészáros LA, Szabó E, *et al.*. Real-time release testing of dissolution based on surrogate models developed by machine learning algorithms using NIR spectra, compression force and particle size distribution as input data. Int J Pharm [Internet]. 2021;597:120338. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0378517321001423.
45. Eshel G, Levy GJ, Mingelgrin U, Singer MJ. Critical evaluation of the use of laser diffraction for particle-size distribution analysis. Soil Sci Soc Am J Wiley. 2004;68:736–43.
46. Rantanen J, Räsänen E, Antikainen O, Mannermaa JP, Yliruusi J. In-line moisture measurement during granulation with a four-wavelength near-infrared sensor: an evaluation of process-related variables and a development of non-linear calibration model. Chemom Intell Lab Syst. 2001;56:51–8.
47. Chen T, Morris J, Martin E. Gaussian process regression for multivariate spectroscopic calibration. Chemom Intell Lab Syst. 2007;87:59–71.
48. Mendyk A, Pacławski A, Szafraniec-Szczęsny J, Antosik A, Jamróz W, Paluch M, *et al.*. Data-driven modeling of the bicalutamide dissolution from powder systems. AAPS PharmSciTech. 2020;21.
49. Salehinejad H, Kitamura J, Ditkofsky N, Lin A, Bharatha A, Suthiphosuwan S, *et al.*. A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography. Sci Rep Nature Publishing Group. 2021;11:1–11.
50. Mowbray M, Vallerio M, Perez-galvan C, Zhang D, Del A, Chanona ADR, *et al.*. Reaction Chemistry & Engineering industries †. React Chem Eng [Internet]. Royal Society of Chemistry; 2022; Available from: https://pubs.rsc.org/en/content/articlepdf/2022/re/d1re00541c.
51. 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. Available from: https://doi.org/10.1208/s12248-022-00725-x.

Authors and Affiliations

Hussain Ali[1] · Prakash Muthudoss[2] · Manikandan Ramalingam[3] · Lakshmi Kanakaraj[3] · Amrit Paudel[4,5] · Gobi Ramasamy[1]

[1] Christ (Deemed to Be University), Bangalore 560029, Karnataka, India

[2] A2Z4.0 Research and Analytics Private Limited, Chennai 600062, Tamilnadu, India

[3] Chettinad School of Pharmaceutical Sciences, Chettinad Academy of Research and Education, Chettinad Health City, 603103 Chennai, Tamilnadu, India

[4] Research Center Pharmaceutical Engineering GmbH (RCPE), Inffeldgasse 13, 8010 Graz, Austria

[5] Institute of Process and Particle Engineering, Graz University of Technology, Inffeldgasse 13/3, 8010 Graz, Austria