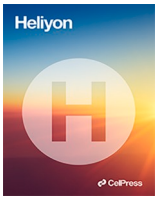
Heliyon 10 (2024) e39838

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Optimization of a tensile strength prediction model for compacted ribbons using NIR-HIS analysis

Juthamat Wanfueangfu [a], Jetsada Posom [b], Duchdoune Teerasukaporn [c], Panuwat Supprung [d], Jomjai Peerapattana [a,*]

[a] Division of Pharmaceutical Technology, Faculty of Pharmaceutical Sciences, Khon Kaen University, Khon Kaen 40002, Thailand
[b] Department of Agricultural Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, 40002, Thailand
[c] Medica Innova Research and Development, Medica Innova Co., Ltd., Bangkok, 10310, Thailand
[d] Department of Postharvest and Agricultural Process Engineering, Faculty of Engineering, Rajamangala University of Technology Isan, Khon Kaen Campus, Khon Kaen 40000, Thailand

ARTICLE INFO

Keywords:
Near infrared hyperspectral imaging spectroscopy
Variable importance in projection
Competitive adaptive reweighted sampling
Genetic algorithm
Roller compactor
Variable selection method

ABSTRACT

The tensile strength (TS) of compacted ribbon is a critical quality attribute in the roller compaction process that impacts the quality of the finished product. This study investigated the use of Near Infrared Hyperspectral Imaging Spectroscopy (NIR-HIS) technology for predicting TS of compacted ribbons, considering the effects of surface curvature, different spectral preprocessing methods, and variable selection methods on a predictive model based on Partial Least Squares regression (PLSr). The spectral preprocessing methods evaluated were Mean Centering (MC) and Standard Normal Variate (SNV). The variable selection methods were Filter by Regression Coefficient (REG), Variable Importance in Projection (VIP), Competitive Adaptive Reweighted Sampling (CARS), and Genetic Algorithm (GA). The results indicated that curved surfaces had no significant impact on the predictive performance of the model (p-value of 0.39 for RMSEP). The PLSr-CARS method, combined with MC spectral preprocessing, was successful in selecting and reducing the number of wavelengths from 182 to 5, as indicated by high values of R^2_{pred} and RPD, and a low RMSEP value (0.97, 5.75, and 7.60 %, respectively). An MLR model using the 5 wavelengths was also developed, showing similar performance to the PLSr model. Both the MLR and PLSr models demonstrated high predictive accuracy and reliability. These models can perform well even when developed using only a few wavelengths, leading to significant reductions in processing time and measurement costs, making them valuable tools for quality control in the pharmaceutical industry.

1. Introduction

Roller compaction is a crucial process in the dry granulation technique, which addresses powder-related issues such as poor flowability, low bulk density, segregation, and dustiness [1]. This method is suitable for heat- and moisture-sensitive drugs as it does not involve solvents or drying steps. During roller compaction, powder is fed into two counter-rotating rolls and compacted, resulting in ribbons. These ribbons are then milled into granules.

* Corresponding author.
E-mail address: jomsuj@kku.ac.th (J. Peerapattana).

https://doi.org/10.1016/j.heliyon.2024.e39838
Received 18 September 2023; Received in revised form 18 October 2024; Accepted 24 October 2024
Available online 25 October 2024

Roller compactors have different operational models, including the fixed-gap roller model, where the distance between roll gaps is constant. However, the force applied to the powder beds can vary depending on the fluctuation of mass fed into the gap. This results in inconsistent compaction force, causing variations in the tensile strength (TS) of the ribbons [2]. These variations can affect the quality of the granules because low TS ribbons are more likely to break down into a fine powder during milling [3]. Such variations can also negatively affect the subsequent blending process, where the powder may not flow as well as larger granules, leading to the loss of improved flow properties and a higher amount of fine particles requiring re-compaction, ultimately affecting the drug content of the granules [4].

As roller compaction is a high-speed, continuous unit operation, it is essential to monitor ribbon TS in real-time and adjust roller parameters accordingly to maintain the quality of the compaction process. The conventional three-point bending test method for determining ribbon TS is destructive, slow, and causes yield loss due to the inability to return samples to the manufacturing line. Measuring the ribbon's width and length takes approximately 5 min, which is too long to adjust roller compactor parameters to maintain ribbon TS consistency. Therefore, a solution to this problem aligns with the concept of Process Analytical Technology (PAT), which focuses on understanding the process to maintain critical quality attributes. Non-destructive techniques such as Near Infrared Spectroscopy (NIRS) and Near Infrared-Hyperspectral Imaging Spectroscopy (NIR-HIS) have been developed to monitor the roller compaction process. These methods are fast and do not harm the product. NIR-HIS provides the added benefit of presenting the distribution of chemical or physical properties, whereas NIRS provides only an average of these properties. By coupling NIR-HIS with chemometric methods, valuable information can be obtained for monitoring the manufacturing process from raw material to finished product and identifying the root cause of manufacturing problems.

Several studies have utilized NIR-HIS to monitor crucial physical attributes of roller compaction, employing different types of roll surfaces. Khorasani et al. employed principal component analysis (PCA) to visualize the porosity distribution in ribbons compacted by smooth roll surfaces [5] and predicted the percentage of porosity using Partial Least Squares regression (PLSr) [6]. Another study by Souihi et al. successfully visualized the density distribution in ribbons compacted by knurled roll surfaces through PCA and PLSr, demonstrating a strong correlation between observed and predicted values for the studied response [7]. One study conducted in a roller compactor with a combination of knurled and smoothed surfaces addressed the impact of the bended ribbon piece on NIR spectrum variation. This study developed a prediction model for ribbon density using a linear regression relationship between the spectrum slope and the envelope density value. The findings indicated that the bended ribbon piece gave lower absorbance values, which may consequently result in poor predictions of the model [8]. In the pharmaceutical manufacturing factory, there is a roll compactor utilizing a combination of smooth and fluted roll surfaces that produces ribbons with one planar and one curved surface. To date, no study that has explored the effects of spectrum variation due to the curved ribbon surface on the prediction performance of the model. If the curvature surface has a significant impact on the prediction performance, it becomes necessary to optimize the in-line measuring position or technique, and additional data treatment techniques may be required. This is particularly important since real-world production lines cannot provide consistent one-sided measurements due to the random movement of ribbon samples.

One of the major challenges of using NIR-HIS for online monitoring of the manufacturing line is the time required to process the full spectrum with spatial information. This involves dealing with large volumes of data, which may contain noise from the instrument as well as uninformative data, leading to complexity and poor predictive models [9]. Therefore, variable selection methods such as filtering by regression coefficient (REG), Variable Importance in Projection (VIP), Competitive Adaptive Reweighted Sampling (CARS), and Genetic Algorithm (GA) can be used to select key informative wavelengths for more accurate and simpler models while reducing the measurement time and cost of the instrument. These variable selection methods have been studied in the pharmaceutical field with both NIRS and NIR-HIS. Ravn et al. utilized VIP to select the wavelengths before generating a PLSr model for a three-component concentration prediction without comparing to the full wavelength [10]. Additionally, Abrahamsson et al. successfully employed GA to reduce the prediction error of Active Pharmaceutical Ingredient (API) concentration by 15 % compared to the full wavelength model [11]. After selecting the key informative wavelengths using variable selection methods, the implementation of a more feasible instrument called Near Infrared Multispectral Imaging Spectroscopy (NIR-MIS) becomes possible. NIR-MIS is designed to operate within a smaller range of wavelengths, specifically targeting the selected informative wavelengths. This narrower range allows for a more compact and cost-effective instrument, making it suitable for installation in production lines.

Therefore, the aims of this study were to (1) investigate the effect of curved surfaces on the prediction accuracy of the TS prediction model by NIR-HIS; (2) select key informative wavelengths using various wavelength selection methods including REG, VIP, CARS, and GA as well as spectral preprocessing techniques; and (3) generate distribution maps of TS of the ribbon sample. The ribbon samples used in this study were manufactured from the same machine and method as the manufacturing site. This often resulted in an inconsistency in the ribbon's TS, which gave rise to a high number of fine particles and required multiple re-compaction cycles.

2. Material and methods

2.1. Materials

The powder mixture was composed of 22.45 % model drug, 58.4 % microcrystalline cellulose PH102 (MCC102), 14.6 % croscarmellose sodium (CCS), 2.2 % colloidal silicon dioxide (CSD), and 2.43 % magnesium stearate (MgSt). All chemicals used were of pharmaceutical grade. The powder mixture was mixed in a lab-scale cube mixer (Erweka, Heusenstamm, Germany). Blend homogeneity was reached after 5 min mixing at 34 rpm.

2.2. Roller compaction

To produce ribbons with varying TS, a roller compactor equipped with two counter-rotating smooth and fluted surface rolls (Fig. 1A and B) was continuously operated. Five kilograms of powder mixture were fed into the feed hopper through the feed screw and then compacted in the compaction zone during the first cycle. It was then milled through sieves No. 12 and 18. The intermediate product was then passed through a high-speed granulator with sieve No. 30. The granules were separated using sieve No. 60, and the fine particles that passed through sieve No.60 were recompacted in a second cycle. The roller distance and main roller speed remained constant at 3 mm and 3 rpm, respectively. The feed screw speed and hydraulic pressure were varied during each compaction cycle (Table 1) to achieve different levels of load on the roller. Compact ribbons, measuring approximately 30*10*3 mm (Fig. 1C and D) were collected. A total of 24 ribbons (from 8x3) obtained from 8 different compaction conditions with three replicates were collected and subsequently stored at 25 °C and 45 % RH.

2.3. NIR hyperspectral image measurement

The experimental setup was demonstrated in Fig. 2. The total of ribbons was 24 samples from 8 different tensile strength levels. Each level had three replicates. Each ribbon was assigned as individual sample. A line-scan NIR-HIS, SISUChema XL TM Chemical Imaging Workstation (Specim, Finland) equipped with ImSpector TM N25E imaging spectrograph (Specim, Finland) was used. White and dark reference measurements were done to calibrate the equipment. Ribbon samples were placed on the stage, moved, and scanned in the range of 900–1700 nm with 5.0 ms exposure, 3.2 nm resolution, and a 10 mm/s conveyor speed. For a total of 24 samples, three ribbons were scanned as representatives of each roller compaction parameter level. The ribbons were scanned on both the planar and curved sides, therefore the total image was 48. Reflectance of samples was collected as three-dimensional data. The X and Y axes indicate spatial or pixel information, while the Z-axis represents wavelength information. The equipment provided 320*803 spatial pixel (X*Y) and 256 wavelengths (Z).

Relative reflectance image is used in hyperspectral image analysis rather than absolute reflectance or raw hyperspectral image because it corrects uneven light intensity and inconsistent voltage or current. This is achieved by comparing the amount of light reflected from a sample to that reflected from white and dark reference surfaces. The white reference is assumed to have a reflectance value of 100 %, and the dark reference is assumed to have a reflectance value of 0 %. Relative reflectance image was calculated as Eq. (1) [12–17].

$$R = \frac{I_{raw} - I_{dark}}{I_{white} - I_{dark}} \tag{1}$$

Where R is the relative reflectance image, I_{raw} is the raw hyperspectral image, I_{white} is the white reference image, and I_{dark} is dark reference image.

Hypercube data of the ribbons were collected. A region of interest (ROI) was selected by removing the background using principal component analysis (PCA). The PC1 matrix, which explained most of the variation in the sample, was reshaped into a 2D image (Fig. 3B). A threshold PC1 score was used to separate the ROI from the background. The ROI image was presented as a binary image, with the yellow part presenting the ROI (Fig. 3C), which is identical to the sample region shown in Fig. 1C. The average spectrum of the ROI of each sample was used as the representative spectrum for generating a prediction model [7,13–17].

2.4. Tensile strength measurement

The TS of the ribbons was quantified by a three-point beam bending test using a texture analyzer (TA. XT Plus, Texture Technologies, Scarsdale, NY, USA) with Texture Expert Exceed (ver. 2.50) software. The length of the gap distance (L) between the lower 2

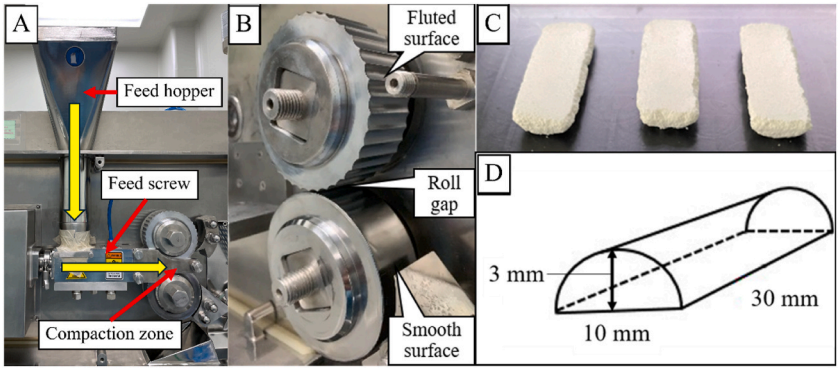


Fig. 1. (A) Roller compactor. (B) smooth and fluted surface roll surface. (C) ribbons compacted in 1st cycle with 150 rpm feed screw speed and 22 MPa hydraulic pressure. (D) diagram of compacted ribbon.

Table 1
Roller compaction parameters.

Cycle	Feed screw (rpm)	Hydraulic pressure (MPa)	Load at roller (kg)
1	100	10	250–400
	150	15	450–550
	150	22	550–700
	150	25	700–800
2	70	4	300–400
	70	5	400–500
	100	10	500–600
	150	15	600–700

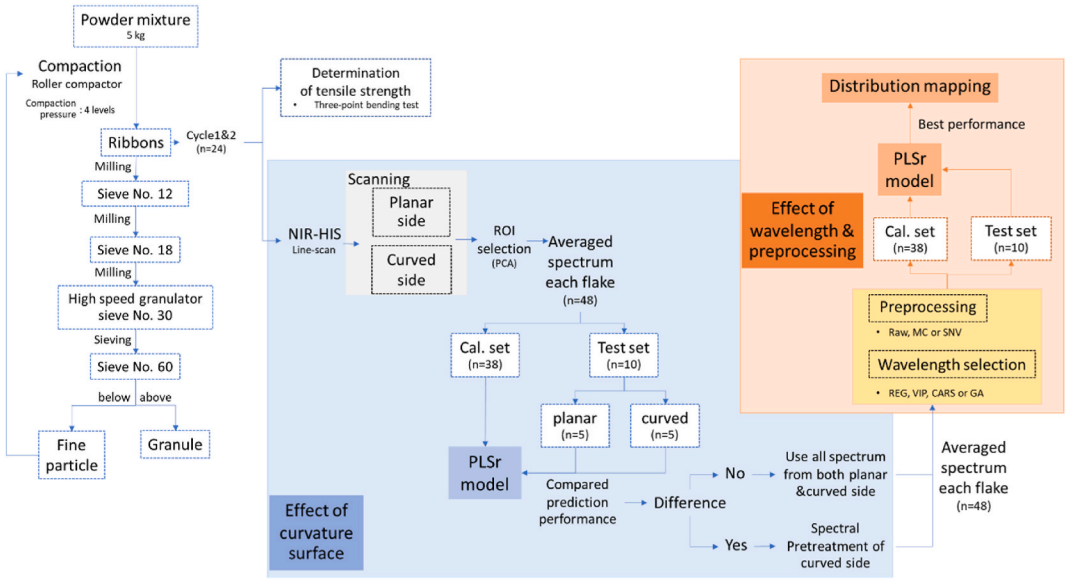


Fig. 2. Schematic representation of the experimental steps used to acquire distribution mapping of compacted ribbon TS. Cal: calibration, ROI: region of interest, PLSr: partial least squares regression.

beams was set at 15.4 mm. The test rate was set at 200 points/s and a test speed of 0.2 mm/s. The instrument was calibrated prior to data collection using a 1 kg weight. Force (F) at break values were collected such that the flat side of the ribbon was always faced towards the two lower stationary beams. Thickness (t) and width (W) of the ribbons were measured using a Digital Caliper. The TS was calculated as Eq. (2) [18,19]:

$$\sigma_T = 3FL/2Wt^2 \tag{2}$$

where σ_T is the TS (Pa); F is the force applied at fracture (kN); W and t are the width (mm) and thickness (mm) of the ribbons, respectively; L is the gap distance between two supporting beams underneath the ribbons (mm). The third beam moves downwards to bend the compact right at the middle of the two supporting beams.

2.5. Modelling and spatial hyperspectral image mapping of analytes

To find the most effective model, the models were developed using different variable selection methods and different spectral preprocessing method. The model's performance developed using full wavelength and various wavelength selection included filter by REG, VIP, CARS, and GA were compared. Additionally, the effectiveness of spectral preprocessing techniques, including Mean Centering (MC) and Standard Normal Variate (SNV), was compared. The effective model established, was used to generate a distribution map of ribbon TS. A schematic representation of the experimental steps used to acquire distribution mapping of compacted ribbon TS is illustrated in Fig. 2.

Step 1 focuses on developing an effective model to predict ribbon TS while investigating the influence of surface curvature on prediction performance. The presence of a curved surface can reduce the intensity of the spectrum, affecting the prediction values. Hence, it is essential to address the variation caused by the curved surface during the prediction process. An investigation into the effect of surface curvature on prediction accuracy was conducted by obtaining 48 spectra. These spectra were obtained by averaging the spectra within each region of interest (ROI) from 24 compacted ribbon samples.

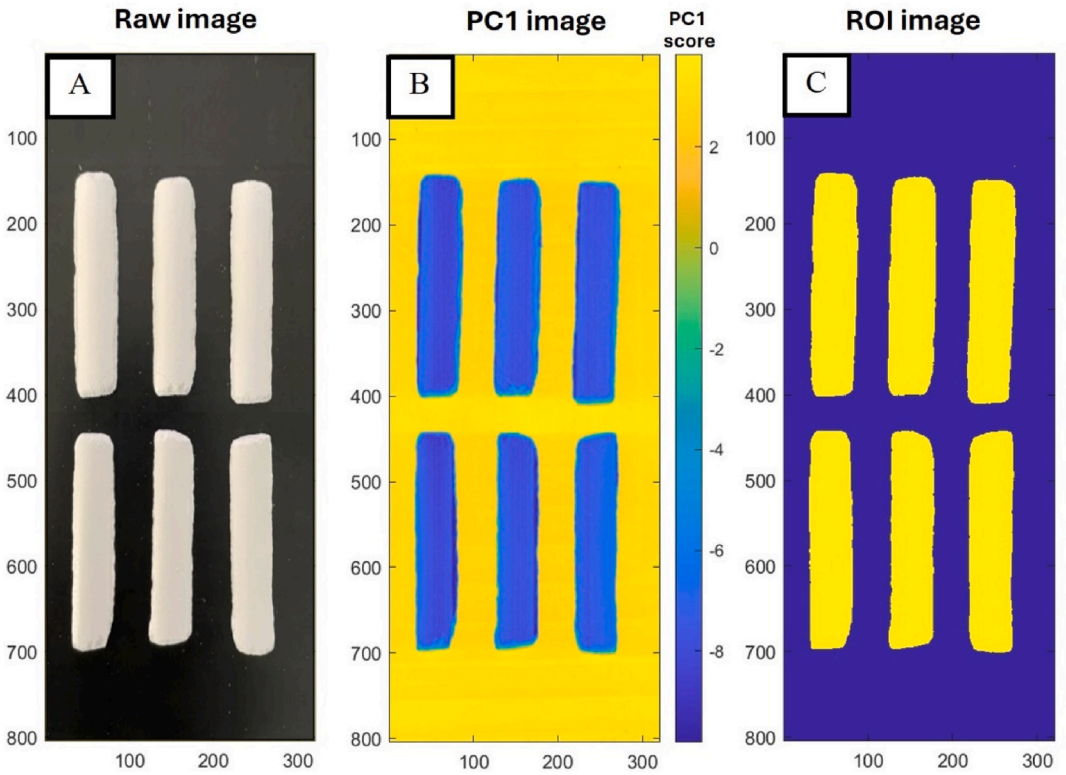


Fig. 3. (A) Raw image, (B) PC1 score distribution image and (C) ROI region presented as binary image.

To reduce prediction performance bias, five subsets were generated, each containing full wavelength spectra data. From the total of 24 samples, each subset includes 5 samples randomly selected to form the test set. Within this test set, there are 5 planar spectra and 5 curvature spectra. The remaining 19 samples, along with the 38 spectra they represent, were assigned to the calibration set.

Partial Least Squares Regression (PLSr) and leave-one-out cross-validation were employed to generate and validate the calibration models for each subset. Subsequently, these models were employed to predict the test set separately for both plane and curve sides. To compare the prediction accuracy, the Root Mean Squared Error of Prediction (RMSEP) for each side (plane and curve) was calculated. Furthermore, T-test analysis was conducted to statistically evaluate the RMSEP between the plane and curve side predictions for each model.

To avoid overfitting, the optimal number of factors was selected based on the highest coefficient of determination (R^2) value, the lowest root mean square error of cross-validation (RMSECV), and the root mean square error of calibration (RMSEC) to RMSECV ratio not being less than 0.8. The PLSr model was then used to predict the test set, and the model with the highest prediction performance was determined based on the highest R^2, lowest root mean square error of prediction (RMSEP) and the highest relative percent difference (RPD). All data analysis was performed using MATLAB (R2022b, 40846673) with in-house code. PLSr was used to reduce the variables into a new dimensional space represented by latent variables or PLS factors, and the PLS equations are presented as Eq. (3) and Eq. (4) [14,20]:

$$X = TP' + E \tag{3}$$

$$Y = Uq' + f \tag{4}$$

Where X is the matrix of spectra (n x p). Y is the response vector (n x 1). T and U (n x h) are the score of X and Y, respectively. P' (h x p) and q' (h x 1) are the loadings of matrix X and vector Y, respectively. E (n x p) and f (n x 1) are matrix and vector of the residual, respectively. The inner relationship of score of X and Y from the least-square method is described as $U = WT$ where W denotes the weight of inner relationship [20].

If the curved surface had a significant impact on the prediction performance, the spectra within the region of interest (ROI) needed to be preprocessed prior to being averaged and selected as representative for that sample. Alternatively, only spectra from planar surfaces were selected to develop the model. If the curved surface had no significant impact on the prediction performance, the spectra from both the planar and curved sides were used to develop the model.

Step 2, to develop an effective prediction model, a suitable number of wavelengths was selected to generate the PLSr model by using wavelength selection models that included REG, VIP, CARS and GA. In addition, proper spectral preprocessing methods were

investigated that included MC and SNV. The most effective model was selected by comparison of the predictive performance include R^2 of prediction (R^2_{pred}), RMSEP and RPD.

REG is a filtration technique that selects the most predictive information of each wavelength presented by the regression coefficient of the PLSr model, which is generated from the full wavelength range. The matrix of the optimal number of wavelengths was selected to build the model.

VIP is also a filtration-based technique. The wavelength with VIP scores more than 1 were selected to generate the model. VIP score can be calculated as Eq. (5) [21,22]:

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{F} w_{jf}^2 \bullet SSY_f \bullet J}{SSY_{total} \bullet F}} \tag{5}$$

The weight value for a given variable j and component f is denoted by w_{if} , while SSY_f represents the sum of squares of the explained variance for component f, considering J number of X variables. SSY_{total} corresponds to the total sum of squares of the explained variance of the dependent variable, while F is the total number of components. VIP_j denotes the importance of the j variable and is a measure of its contribution to the variance explained by each PLS component, presented as w_{jf}^2.

CARS is a sequential forward selection algorithm that uses a Monte Carlo sampling method to select wavelengths. At each step, CARS selects the variable that provides the greatest improvement in model fit, based on Root Mean Square Error of Cross Validation (RMSECV), while competing against other variables already selected. The algorithm also uses an adaptive reweighting scheme to balance the selection of variables that are highly correlated with each other. By iteratively selecting and eliminating variables, CARS aims to obtain a parsimonious model that achieves good prediction accuracy and generalizability [23].

In GA, a genetic algorithm is used to select the most important predictor variables for a PLSr model. The algorithm starts by randomly generating a set of candidate solutions, each of which represents a subset of predictor variables. The fitness of each candidate solution is evaluated based on the prediction performance of a PLSr model using the selected variables. The fittest solutions are selected for reproduction, and crossover and mutation operations are applied to create new candidate solutions. This process continues for a specified number of generations, with the goal of improving the prediction performance of the PLSr model by selecting the most informative predictor variables [24].

In the case of CARS and GA, the subset of selected variables differed between each sampling run. The method was then performed ten times to increase confidence in the selected variable subset. Subsequently, the subset with the lowest number of wavelengths and the highest prediction performance was chosen.

For the spectral preprocessing, either raw spectra or preprocessed spectra were applied for model development to compare their accuracy. Spectral preprocessing holds promise in enhancing the performance of the calibration model by effectively eliminating noise and background interferences, thereby leading to improved accuracy and reliability. Spectra were pretreated with mean centering (MC) and standard normal variate (SNV).

MC involves subtracting the average value of each spectrum from each data point in that spectrum using a simple equation: $Xmc_{ij} = X_{ij} - \overline{X_{ij}}$ where Xmc_{ij} and X_{ij} is the matrix in row i and column j of the MC and original spectrum, respectively $\overline{X_{ij}}$ is mean of the original spectrum. This process can help to remove variation in the data that is not informative for the model and can lead to better predictions.

SNV is useful for removing baseline offsets and correcting for differences in sample path length and intensity of the measured spectra using basic equation: $Xsnv_{ij} = \left(X_{ij} - \overline{X_{ij}}\right)/SD$ where $Xsnv_{ij}$ and X_{ij} is the matrix in row i and column j of the SNV and original spectrum, respectively. $\overline{X_{ij}}$ and SD is mean and standard deviation of the original spectrum, respectively.

2.6. Distribution mapping

A distribution map of ribbon TS is a visual representation or map that illustrates the spatial distribution or pattern of TS across a region in the ribbon sample. It helps in understanding the pattern of TS distribution and enables monitoring and visualization of the data. The optimal TS calibration model was implemented to predict TS (Y_{pred}) from the spectrum of sample in each pixel by equation: $Y_{pred} = XB$ where X denotes the relative reflectance in each pixel. B denotes the regression coefficient of calibration model. Y_{pred} were then present in the 2D color image while scale of color presents the relative TS. The average TS of all pixels can be used to estimate an

Table 2
Reference value of TS of compacted ribbons.

Cycle	Load at roller (kg)	TS (MPa)				RSD (%)
		No.1	No.2	No.3	Mean	
1	250–400	0.73	0.82	0.80	0.78	6.02
	450–550	0.96	1.01	1.20	1.06	11.83
	550–700	1.65	0.58	0.88	1.04	53.00
	700–800	0.84	1.53	2.13	1.50	43.04
2	300–400	1.02	1.06	1.02	1.04	2.05
	400–500	1.21	0.96	1.01	1.06	12.60
	500–600	1.85	1.76	1.46	1.69	12.14
	600–700	2.29	2.57	2.26	2.38	7.22

average TS of sample.

3. Results and discussion

3.1. TS determination

The reference TS value of compacted ribbon was determined using a three-point bending test, and the results are presented in Table 2. The mean TS value ranged from 0.78 to 2.38 MPa, and the results showed a high Relative Standard Deviation (RSD) (2.05–53.00 %) in each sample, indicating fluctuations in TS within the same roller compactor parameters. Consequently, TS value and NIR spectrum were correlated individually for modeling.

3.2. Effect of ribbon surface curvature

The impact of curvature on the accuracy of TS prediction using optical imaging has been widely recognized as a multi-scattering effect [25] and a significant source of error [26,27] To assess this effect, the TS prediction performance, as indicated by R^2_{pred} and RMSEP values, was compared for planar and curved surfaces. To this end, five PLSr models were developed using a calibration set containing 38 spectra randomly selected from both planar and curved surfaces. Each model was then used to predict two test sets, one composed of planar surface spectra and the other of curved surface spectra. The results, as shown in Table 3, indicate that the mean R^2_{pred} and RMSEP values for planar surfaces were 0.92 and 10.50 %, respectively, while for curved surfaces, they were 0.93 and 9.45 %, respectively. However, the difference in prediction performance between the two surfaces was not statistically significant, as indicated by a p-value of 0.39 for RMSEP. These results imply that while the multi-scattering effect did occur, the variation resulting from this effect was included in the PLSr model. As a result, the model generated from spectra of both planar and curved surfaces did not require any spectral correction techniques for the curved surface spectra prior to their inclusion in the model.

3.3. TS prediction

To obtain the most effective TS prediction model, two factors were evaluated: the range or number of wavelengths and the spectral preprocessing method. PLSr models were developed using full or partial wavelengths and raw or pretreated spectra with MC or SNV preprocessing methods. The accuracy and fitness of the PLSr models were then compared. The PLSr results are shown in Table 4. The model generated from 10 selected wavelengths with the GA method and a nontreated spectrum gave the best fit for the calibration model, represented by the lowest RMSECV value (8.41) and the highest R^2 values for calibration (R^2_{cal}) and validation (R^2_{val}) (0.97 and 0.96, respectively) with seven latent variables (LVs). The best prediction performance was obtained from the model generated from the 16 selected wavelengths using the GA method with the SNV preprocessing method. The R^2_{pred}, RPD, and RMSEP were 0.97, 6.72, and 6.77 %, respectively with seven LVs.

However, considering the feasibility of implementing the model on the manufacturing line, it is essential to prioritize the lower cost of the instrument by reducing the number of wavelengths for model development. Therefore, a more suitable model would be to focus on achieving a balance between cost-effectiveness and model performance. The PLSr results revealed that only five selected wavelengths from the CARS method could provide an effective model.

The CARS process was applied to extract important wavelengths. This process is illustrated in Fig. 4, which depicts the variation in regression coefficients of the 182 wavelength variables during Monte Carlo sampling. Initially, as shown in Fig. 4A, the number of variables decreased rapidly before slowing down. Fig. 4B demonstrates that as the number of samplings increased, the RMSECV value gradually decreased, reaching its lowest point at the 40th sampling. This indicates that during the first 1–39 operations, uninformative variables were removed. However, beyond the 40th sampling, the RMSECV rose sharply, suggesting that important variables were being excluded from the sampling subset, thus worsening model performance. Consequently, the variable subset obtained at the 40th sampling was identified as the optimal important wavelength subset, consisting of the selected variables at 1113, 1132, 1290, 1307, and 1482 nm. This approach resulted in a significant reduction in the number of wavelengths by 97.25 %, from 182 down to 5.

The resulting five selected wavelengths with MC preprocessing are shown in Fig. 5. The color scale corresponds to the value of TS, with red indicating a high TS value and blue indicating a low TS value. At 1113, 1132, 1290, and 1307 nm, samples with higher TS showed an upward shift in spectral intensity, while at 1482 nm, samples with higher TS showed a downward shift in spectral intensity.

Table 3
Comparison of chemometric statistics and prediction performance of PLSr model from spectrum between planar and curved Surfaces.

Run	LVs	Calibration (n = 38)				Prediction (n = 10)		
		R^2_{cal}	R^2_{val}	RMSEC (%)	RMSECV (%)	R^2_{pred}	Planar side (n = 5)	Curved side (n = 5)
							RMSEP (%)	RMSEP (%)
1	5	0.95	0.93	7.34	8.48	0.92	10.08	12.79
2	5	0.94	0.92	10.47	12.17	0.95	8.27	5.97
3	5	0.95	0.93	8.74	10.16	0.92	10.00	8.41
4	5	0.95	0.93	11.13	12.95	0.91	9.72	9.42
5	5	0.95	0.93	8.95	10.51	0.92	14.42	10.66

Table 4
The overview of wavelength selection methods with preprocessing methods and their calibration and prediction performance.

Selection method	Spectral pretreatment	No. of wavelength	Calibration (n = 38)						Prediction (n = 10)		
			LVs	R^2_{cal}	R^2_{val}	Explain	RMSEC (%)	RMSECV (%)	R^2_{pred}	RPD	RMSEP (%)
Full	Raw	182	6	0.95	0.93	100	8.91	10.73	0.95	4.82	9.11
	MC	182	5	0.95	0.92	97.66	8.92	11.02	0.95	4.94	8.88
	SNV	182	5	0.95	0.92	97.76	8.92	11.03	0.95	4.89	8.95
REG	Raw	20	5	0.89	0.85	100	13.22	15.28	0.87	2.83	15.16
	MC	9	4	0.92	0.89	97.19	11.32	12.9	0.92	3.59	12.03
	SNV	7	5	0.94	0.92	97.39	9.69	11.35	0.95	5.38	9.15
VIP	Raw	21	5	0.92	0.88	98.06	11.24	13.69	0.92	3.63	12.19
	MC	25	6	0.95	0.92	99.42	8.92	10.97	0.93	4.09	11.23
	SNV	22	6	0.95	0.93	99.49	9.08	10.73	0.93	4.08	11.08
CARS	Raw	13	6	0.96	0.95	100	7.43	9.14	0.95	4.87	9.05
	MC	***5***	***5***	***0.96***	***0.95***	***97.77***	***7.57***	***8.72***	***0.97***	***5.75***	***7.6***
	SNV	8	5	0.95	0.94	97.76	8.44	9.88	0.96	5.26	8.55
GA	Raw	10	7	0.97	0.96	100	6.99	8.41	0.95	4.33	9.61
	MC	13	5	0.96	0.95	98.13	7.78	9.18	0.96	5.08	8.64
	SNV	16	7	0.96	0.94	98.36	7.65	9.52	0.97	6.72	6.77

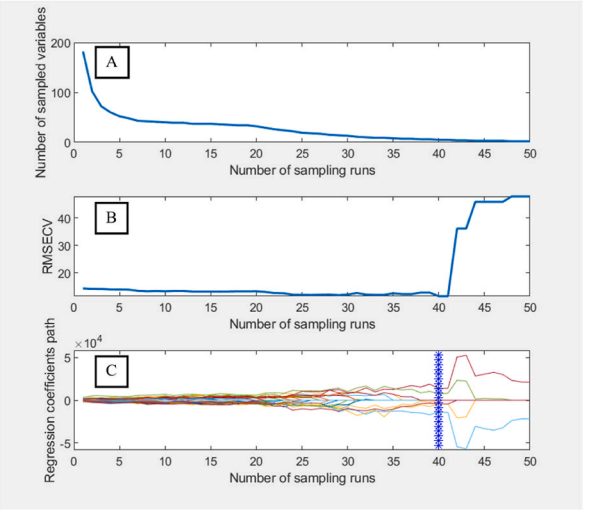


Fig. 4. (A)Variations in the number of sampled variables, (B) RMSECV, and (C) the regression coefficient paths were observed with incresing of number of sampling run.

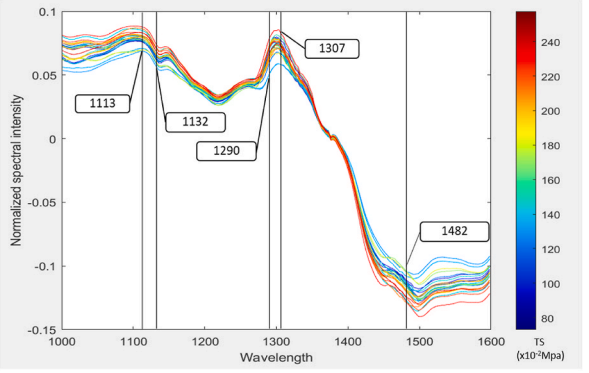


Fig. 5. The five selected wavelengths in spectral with mean centering. Red indicating a high value of TS and blue indicating a low value of TS.

The selected wavelengths for the model correspond to various stretching vibrations and overtone bands, reflecting the chemical composition and bonding in the tablet's components, as described below [28,29]. The wavelengths at 1113 nm and 1132 nm are associated with the second overtone of C-H stretching vibrations, linked to C-H bonds in the aromatic groups of the model drug and alkyl groups (CH_3) of the model drug, MCC102, and CCS. The wavelengths at 1290, 1307, and 1482 nm are in the first overtone region of water O-H stretching vibrations [29,30]. These wavelengths are associated with moisture content and components such as MCC102 and CCS. MCC102, which constitutes 58.4 % of the tablet composition, plays a crucial role in tablet strength through strong hydrogen bonding [31]. Additionally, the wavelength at 1482 nm is also in the region associated with the first overtone of N-H stretching vibrations of NH_2, which can be found in the model drug.

The PLSr model resulted in five sufficient latent variables (LVs), explaining 97.77 % of the variation. The calibration model parameter values were 0.96 (R^2_{cal}), 0.95 (R^2_{val}), 7.57 % (RMSEC), and 8.72 % (RMSECV). The ratio of RMSEC to RMSECV was 0.901, which is close to 1, indicating a good fit of the calibration model. The prediction performance of the selected model was also high, with R^2_{pred}, RPD, and RMSEP values of 0.97, 5.75, and 7.60 %, respectively. Although the prediction accuracy dropped compared to the model developed from the full spectrum, it was still higher (with LVs, R^2_{pred}, RPD, and RMSEP values of 6, 0.95, 4.82, and 9.11 %, respectively). Moreover, the validation parameters of the selected model fall within the acceptable criteria for use in quality assurance. If $0.92 < R^2 < 0.96$ and RPD >5, the model is considered to have excellent prediction accuracy and is suitable for quality monitoring [32]. Therefore, this optimal model was used to generate the distribution mapping of TS in the ribbon.

The primary prediction model used was PLS. However, concerns arose due to the few variables used to generate the model. The

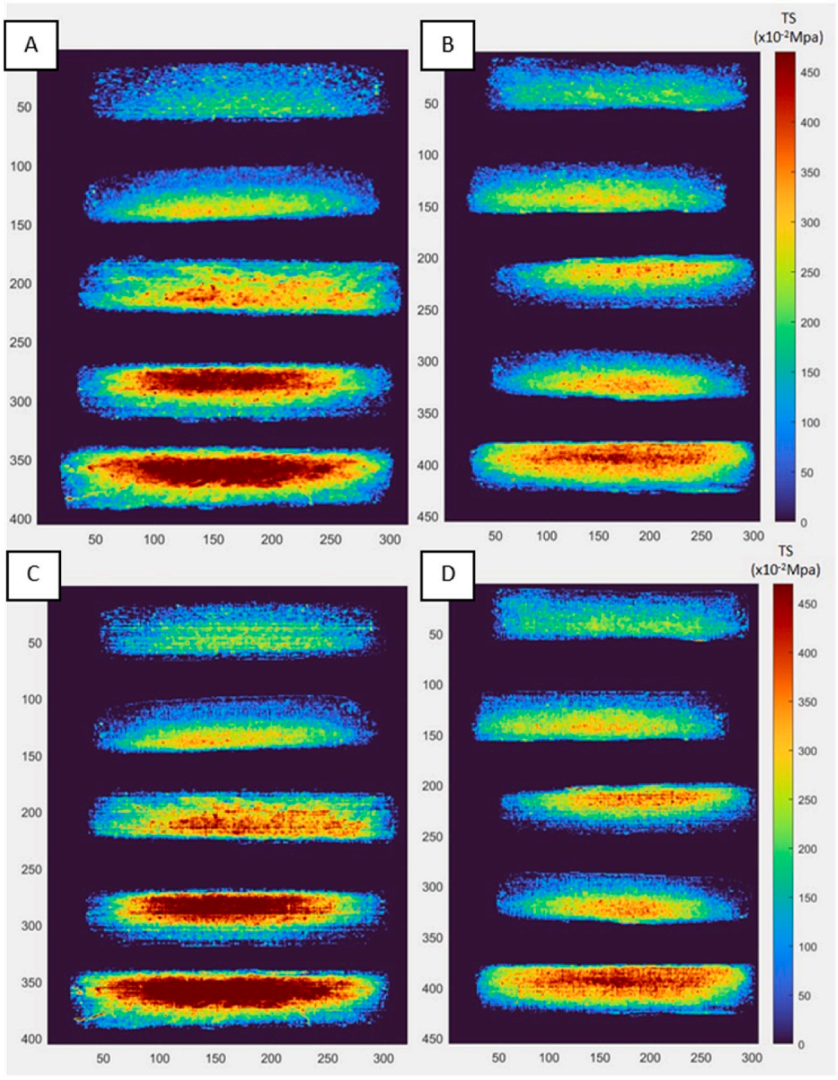


Fig. 6. Distribution maps of TS along the ribbon. (A, C) the planar side predicted from full spectrum and five spectra, respectively which have 0.73, 1.53, 1.02, 1.85, and 2.57 MPa of TS reference value. (B, D) the curve side predicted from full spectrum and five spectra, respectively which have 0.88, 1.20, 0.96, 1.01 and 2.26 MPa of TS reference value.

model was generated using five selected wavelengths, and the PLS model utilized five latent variables. This implies that the model does not reduce the dimension of the variables as intended by PLS. Consequently, we decided to compare it to the MLR method, which does not reduce variable dimensions. All variables were used to find the correlation between the independent variables (X_i) and the dependent variable (Y) as expressed by the followed equation (Eq. (6)).

$$Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n + e \tag{6}$$

where a_i are regression coefficients, n is the number of variables and e is the model constant or residual [33,34]. Consequently, a straightforward MLR model was directly established using the same five selected wavelengths. The linear function is presented by the following equation (Eq. (7)).

$$Y_{TS} = 1420.07X_{1113} + 14549.05X_{1132} + 7468.55X_{1290} + 3877.56X_{1307} + 4355.67X_{1482} + 5.29 \tag{7}$$

where Y_{TS} denotes the predicted TS value and X_i denotes the relative reflectance value at the wavelength of i nm.

This MLR model exhibited excellent calibration performance, with R^2_{cal}, R^2_{val}, RMSEC and RMSECV of 0.96, 0.96, 8.27 % and 8.84 %, respectively. The prediction performance was also remarkable, with R^2_{pred}, RPD and RMSEP values of 0.97, 5.73 and 7.73 %. respectively. These results were not different from those of the PLSr model and fall within the acceptable criteria previously mentioned for use in quality assurance. Therefore, both equations are suitable for generating the distribution map. MLR provides a prediction model with less complexity through an uncomplicated equation. However, after checking the correlation among the five selected wavelengths, all parameters showed multicollinearity with $R > 0.9$. Therefore, The PLS model which could be lower the multicollinearity effect [20] for generating the distribution map was selected.

In each model developed from the selected wavelengths, the model with the raw spectrum showed lower R^2_{pred} (0.87–0.96) and RPD (2.83–5.45) with higher RMSEP (8.08–15.16 %). The use of MC and SNV spectrum preprocessing achieved an acceptable range of R^2_{pred} with higher RPD (3.59–5.75 and 4.74 to 6.72 for MC and SNV, respectively) and lower RMSEP (7.60 %–12.03 % and 6.77 %–11.08 % for MC and SNV, respectively). These preprocessing methods can reduce the variation that is uninformative for TS prediction, which may occur due to noise and background interference.

3.4. Distribution mapping of TS

The test set consisted of ten hyperspectral cubes representing compacted ribbons with planar and curved surfaces. These cubes covered a range of TS values from 0.73 to 2.57 MPa. The distribution maps were generated and used to evaluate the reliability of the selected optimal model. The model predicted TS in each pixel of the hyperspectral cube, resulting in distribution maps that depicted the relative TS distribution along the ribbon. In the generated distribution maps, dark red coloration indicates areas of high TS, while dark blue coloration indicates areas of low TS. The map images predicted from the spectra of both the planar and curved sides showed the same distribution pattern, and the distribution image generated from either the full or partial wavelength sets showed a non-different pattern. The images in Fig. 6A-D reveal the presence of heterogeneous TS regions on the roller-compacted ribbon with one convex and one flat surface. In the middle of each ribbon, the TS value was high and gradually decreased towards the edges of the ribbon. This result corresponds with prior studies that showed the stress distribution along the compacted ribbon was parabolic and the variation correlated with density variation [35]. Furthermore, it was observed that at the edge of the ribbon, the prediction results presented zero or negative values, indicating some errors in the hyperspectral imaging. This could be due to shading effects, as the light source is from the side. When the light hits the curved shape of the sample, shading occurs at the edges. However, the average tensile strength (TS) prediction was still close to the reference value. To overcome this problem, we suggest further investigating the light source position that can reduce the effect of shading. Unfortunately, this was beyond our reach due to the limitations of equipment with a fixed light source position.

Notably, ribbons produced at higher compression pressures resulted in a higher heterogeneity of TS values compared to those produced at lower pressures. This result is similar to that of Khorasani et al., who predicted the porosity of the compacted ribbon and found that higher heterogeneity is associated with higher compression pressure [6]. Although the parameters studied were different, there is a correlation between the porosity and TS of the compacted ribbon. Porosity is inversely related to the solid fraction, while TS has a positive correlation with solid fraction [36]. According to Guigon and Simon's research, the uneven distribution of TS on ribbons formed by roller compaction is caused by a feeding process that is not consistently compacted [37].

To test the reliability of each PLSr-model, test sets of ribbon sections including spectra from both planar and curved sides ($n = 10$) were used. Each ribbon within the test sets was measured by NIR-HIS, and the TS result was calculated from the average predicted TS in each pixel. The average predicted values were plotted against the TS measured by the reference method using a simple linear regression model. The prediction curves (Fig. 7) showed a strong correlation with $R^2 = 0.97$, and the comparison of prediction and reference values showed no significant difference (p-value = 0.26). Therefore, NIR-HIS is capable of characterizing differences in TS as a function of position on the ribbon and predicting the TS of the roller-compacted ribbon.

This approach can be utilized to create a distribution map of tensile strength. By reducing the spectral data to five selected wavelengths, it is more practical to develop cost-effective NIR-MIS equipment. This reduction in wavelength coverage not only lowers the price of the equipment but also shortens the measurement time, making it more accessible for pharmaceutical companies.

One of the most significant applications of this technology is in controlling the roller compaction process. Roller compaction is influenced by numerous factors, such as the fluctuation of mass flow through the compaction zone, which can result in flakes with uneven tensile strength. This issue is particularly prevalent in fixed-gap roller compaction models, leads to inconsistent compaction

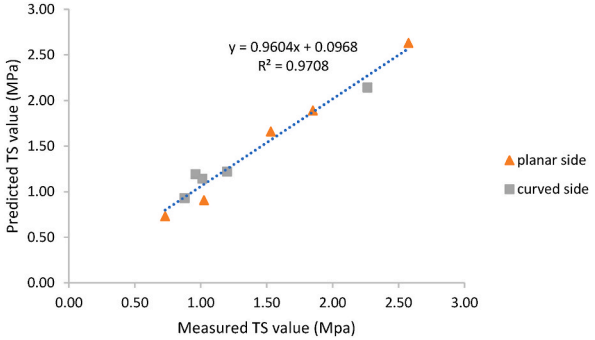


Fig. 7. Correlation of the predicted TS and the reference TS value.

pressure. Currently, the process parameters in roller compaction cannot be completely controlled. Some parameters need to be adjusted at the production line based on the properties of the compacted ribbons. Traditionally, samples of the compacted ribbons are tested for hardness, and process parameters are then adjusted accordingly. However, this method is not efficient, as it involves sampling, testing, and subsequent machine adjustment, which leads to yield loss and delays in real-time adjustments. Moreover, not all compacted ribbons can be sampled, resulting in some products not meeting the required tensile strength criteria.

NIR-MIS offers a solution to these challenges by enabling 100 % visual inspection of samples through a non-destructive method. This allows for real-time adjustments to the process parameters without any yield loss, thus enhancing the efficiency and consistency of the roller compaction process.

4. Conclusion

This study utilized NIR-HIS technology to predict the TS of compacted ribbons. The effect of surface curvature on the prediction performance was examined and, while a multi-scattering effect was observed, it did not significantly impact the predictive performance. An optimal model for predicting the TS of compacted ribbons could be developed using only five wavelengths selected by the CARS method with mean-centered spectral preprocessing. The resulting model exhibited excellent predictive accuracy and reliability. Furthermore, the roller compactor process can be controlled by adjusting the parameters in real-time to acquire consistent compacted ribbon quality. This model is suitable for in-process control of the pharmaceutical dry granulation process by roller compactor within the same powder component. However, variations in chemical bonding can affect the shape of the spectra and may result in errors. For other compositions, the prediction model must be further developed and validated before use. These findings highlight the potential of NIR-HIS technology for application in manufacturing lines. Since the curvature did not affect the model's predictive performance, there is no need for concern with the measurement side. However, the performance of the distribution map can be improved by further studying the light source position to reduce the error occurring at the edge of the sample. Additionally, measuring time and equipment costs can be drastically reduced by selecting a smaller number of wavelengths, and NIR-MIS can be used instead of NIR-HIS.

CRediT authorship contribution statement

Juthamat Wanfueangfu: Writing – review & editing, Writing – original draft, Software, Methodology, Data curation. **Jetsada Posom:** Writing – review & editing, Software, Methodology, Conceptualization. **Duchdoune Teerasukaporn:** Resources. **Panuwat Supprung:** Resources. **Jomjai Peerapattana:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Data availability statement

The study's data has not been uploaded to a publicly accessible repository, but certain data can be provided upon request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to acknowledge the generous support provided by T.O. Chemicals (1979). This research was partially funded by the Graduate School, Khon Kaen University (GSKKU) [641H111] and Faculty of Pharmaceutical Sciences, Khon Kaen University (KKU), Thailand. We would like to acknowledge Dr. Glenn Neville Borlace, for editing the MS via Publication Clinic KKU,

Thailand.

References

[1] B. Olaleye, C.Y. Wu, L.X. Liu, The effects of screw-to-roll speed ratio on ribbon porosity during roll compaction, Int. J. Pharm. 588 (August) (2020) 119770, https://doi.org/10.1016/j.ijpharm.2020.119770.

[2] S.M.A. Alli, Advances in roller compaction/dry-granulation, Res J Pharm Biol Chem Sci. 5 (3) (2014) 1972–1986.

[3] J.M. Rowe, S.T. Charlton, R.J. McCann, Development, scale-up, and optimization of process parameters: roller compaction theory and practice, in: Developing Solid Oral Dosage Forms, second ed., Elsevier, 2017, pp. 869–915, https://doi.org/10.1016/B978-0-12-802447-8.00032-7.

[4] K.M. Hwang, S.Y. Kim, T.T. Nguyen, C.H. Cho, E.S. Park, Use of roller compaction and fines recycling process in the preparation of erlotinib hydrochloride tablets, Eur J Pharm Sci 131 (February) (2019) 99–110, https://doi.org/10.1016/j.ejps.2019.01.036.

[5] M. Khorasani, J.M. Amigo, C.C. Sun, P. Bertelsen, J. Rantanen, Near-infrared chemical imaging (NIR-CI) as a process monitoring solution for a production line of roll compaction and tableting, Eur. J. Pharm. Biopharm. 93 (April) (2015) 293–302, https://doi.org/10.1016/j.ejpb.2015.04.008.

[6] M. Khorasani, J.M. Amigo, J. Sonnergaard, P. Olsen, P. Bertelsen, J. Rantanen, Visualization and prediction of porosity in roller compacted ribbons with near-infrared chemical imaging (NIR-CI), J. Pharm. Biomed. Anal. 109 (2015) 11–17, https://doi.org/10.1016/j.jpba.2015.02.008.

[7] N. Souihi, D. Nilsson, M. Josefson, J. Trygg, Near-infrared chemical imaging (NIR-CI) on roll compacted ribbons and tablets - multivariate mapping of physical and chemical properties, Int. J. Pharm. 483 (1–2) (2015) 200–211, https://doi.org/10.1016/j.ijpharm.2015.02.006.

[8] M.E. Crowley, A. Hegarty, M.A.P. McAuliffe, G.E. O'Mahony, L. Kiernan, K. Hayes, et al., Near-infrared monitoring of roller compacted ribbon density: investigating sources of variation contributing to noisy spectral data, Eur. J. Pharm Sci. 102 (2017) 103–114, https://doi.org/10.1016/j.ejps.2017.02.024.

[9] H. Pu, Selection of feature wavelengths for developing multispectral imaging systems for quality, safety and authenticity of muscle foods-a review, Trends Food Sci. Technol. 45 (1) (2015) 86–104, https://doi.org/10.1016/j.tifs.2015.05.006.

[10] C. Ravn, E. Skibsted, R. Bro, Near-infrared chemical imaging (NIR-CI) on pharmaceutical solid dosage forms-Comparing common calibration approaches, J. Pharm. Biomed. Anal. 48 (3) (2008) 554–561.

[11] C. Abrahamsson, J. Johansson, A. Sparén, F. Lindgren, Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets, Chemom Intell Lab Syst. 69 (1–2) (2003) 3–12.

[12] Global analytical and measuring instruments, Solid sample reflectance measurements. https://www.shimadzu.com/an/uv/support/fundamentals/reflectance_measurements.html, 2020. (Accessed 18 May 2024).

[13] L. Pitak, K. Saengprachatanarug, K. Laloon, J. Posom, Predicting the true density of commercial biomass pellets using near-infrared hyperspectral imaging, Artif. Intell. Agric. 6 (2022) 266–275.

[14] L. Pitak, P. Sirisomboon, K. Saengprachatanarug, S. Wongpichet, J. Posom, Rapid elemental composition measurement of commercial pellets using line-scan hyperspectral imaging analysis, Energy 220 (2021) 119698, https://doi.org/10.1016/j.energy.2020.119698.

[15] K.Q. Yu, Y.R. Zhao, X.L. Li, Y.N. Shao, F. Liu, Y. He, Hyperspectral imaging for mapping of total nitrogen spatial distribution in pepper plant, PLoS One 9 (12) (2014) 1–19.

[16] Y. Shao, Y. Shi, Y. Qin, G. Xuan, J. Li, Q. Li, et al., A new quantitative index for the assessment of tomato quality using Vis-NIR hyperspectral imaging, Food Chem. 386 (November) (2022) 132864, https://doi.org/10.1016/j.foodchem.2022.132864.

[17] K. Yu, Y. Zhao, X. Li, Y. He, NIR hyperspectral imaging for mapping of moisture content distribution in tea buds during dehydration, Int. Agric. Eng. J. 24 (3) (2015) 110–118.

[18] A.V. Zinchuk, M.P. Mullarney, B.C. Hancock, Simulation of roller compaction using a laboratory scale compaction simulator, Int J Pharm 269 (2) (2004) 403–415.

[19] N.P. Davies, M.J. Newton, Mechanical strength, in: G. Alderborn, C. Nystrom (Eds.), Pharmaceutical Powder Compaction Technology, vol. 71, CRC Press, Dekker (NY), 1996, pp. 165–192.

[20] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom Intell Lab Syst. 58 (2) (2001) 109–130.

[21] S. Wold, A. Johansson, M. Cochi (Eds.), PLS-Partial Least Squares Projections to Latent Structures, Escom Science Publishers, Leiden (NL), 1993, pp. 523–550.

[22] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. Chemom. 29 (10) (2015) 528–536.

[23] B.C. Deng, Y.H. Yun, D.S. Cao, Y.L. Yin, W.T. Wang, H.M. Lu, et al., A bootstrapping soft shrinkage approach for variable selection in chemical modeling, Anal. Chim. Acta 908 (2016) 63–74, https://doi.org/10.1016/j.aca.2016.01.001.

[24] K. Hasegawa, Y. Miyashita, K. Funatsu, GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists, J. Chem. Inf. Comput. Sci. 37 (2) (1997) 306–310, https://doi.org/10.1021/ci960047x.

[25] N. Al Makdessi, M. Ecarnot, P. Roumet, G. Rabatel, A spectral correction method for multi-scattering effects in close range hyperspectral imagery of vegetation scenes: application to nitrogen content assessment in wheat, Precis. Agric. 20 (2) (2019) 237–259, https://doi.org/10.1007/s11119-018-9613-2.

[26] J.M. Kainerstorfer, F. Amyot, M. Ehler, M. Hassan, S.G. Demos, V. Chernomordik, et al., Direct curvature correction for noncontact imaging modalities applied to multispectral imaging, J. Biomed. Opt. 15 (4) (2010) 046013.

[27] L. Rogelj, U. Simončič, T. Tomanič, M. Jezeršek, U. Pavlovčič, J. Stergar, et al., Effect of curvature correction on parameters extracted from hyperspectral images, J. Biomed. Opt. 26 (9) (2021).

[28] D.A. Burns, E.W. Ciurczak (Eds.), Handbook of Near-Infrared Analysis, third ed., Taylor & Francis, Dekker (NY), 2008.

[29] H.W. Siesler, Y. Ozaki, S. Kawata, H.M. Heise, Near-Infrared Spectroscopy: Principles, Instruments, Applications. Weinheim (GER), Wiley-VCH, 2002.

[30] Y. Ozaki, T. Genkawa, Y. Futami, Near-infrared spectroscopy, in: Encyclopedia of Spectroscopy and Spectrometry, Elsevier, Amsterdam (NL), 2016.

[31] B.G.E. Reier, R.I.F. Shangraw, Microcrystalline cellulose in tableting, J Pharm Sci 55 (1966) 510–514.

[32] G. Ding, Y. Hou, J. Peng, Y. Shen, M. Jiang, G. Bai, On-line near-infrared spectroscopy optimizing and monitoring biotransformation process of γ -aminobutyric acid, J. Pharm. Anal. 6 (3) (2016) 171–178, https://doi.org/10.1016/j.jpha.2016.02.001.

[33] O.R. Omokungbe, O.G. Fawole, O.K. Owoade, O.A.M. Popoola, R.L. Jones, F.S. Olise, et al., Analysis of the variability of airborne particulate matter with prevailing meteorological conditions across a semi-urban environment using a network of low-cost air quality sensors, Heliyon 6 (6) (2020) e04207, https://doi.org/10.1016/j.heliyon.2020.e04207.

[34] G.K. Uyanık, N. Güler, A study on multiple linear regression analysis, Procedia - Soc Behav Sci. 106 (2013) 234–240.

[35] T. Lecompte, P. Doremus, G. Thomas, L. Perier-Camby, J.C. Le Thiesse, J.C. Masteau, et al., Dry granulation of organic powders - dependence of pressure 2D-distribution on different process parameters, Chem. Eng. Sci. 60 (14) (2005) 3933–3940.

[36] C.K. Chang, F.A. Alvarez-Nunez, J.V. Rinella, L.E. Magnusson, K. Sueda, Roller compaction, granulation and capsule product dissolution of drug formulations containing a lactose or mannitol filler, starch, and talc, AAPS PharmSciTech 9 (2) (2008) 597–604.

[37] P. Guigon, O. Simon, Roll press design - influence of force feed systems on compaction, Powder Technol. 130 (1–3) (2003) 41–48.