International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No. 1 (Jan-24)

http://dx.doi.org/10.12785/ijcds/150107

Introduction of LSSVR for the Prediction of the Yellowness Index

Wan Sieng Yeo[1,2] and Agus Saptoro[1,2]

[1]*Department of Chemical and Energy Engineering, Curtin University Sarawak Campus, CDT 250, Miri 98009, Sarawak, Malaysia*
[2]*Curtin Malaysia Research Institute, Curtin University, Malaysia*

Received 8 May. 2023, Revised 20 Oct. 2023, Accepted 16 Nov. 2023, Published 1 Jan. 2024

Abstract: Data-driven models including principal component regression (PCR), partial least square regression (PLSR), and least square support vector regression (LSSVR) have been widely applied as predictive models in various applications. However, studies employing regression models to estimate the yellowness index (YI) are scarce in the literature. This study, therefore, focuses on developing non-destructive YI measurements using regression models. The collected RGB calculated XYZ and obtained CIE LAB values were set as the input variables. Meanwhile, the YI value was denoted as the output variable. Results indicated that the LSSVR model outperforms PCR and PLSR models in predicting YI in which the root means square errors of LSSVR for the training and testing datasets were found to be 261,406% to 294,218% and 725% to 772% lower than PLSR and PCR, respectively. LSSVR is also attributed to higher coefficients of determination ($\mathbb{R}^2$) that are superior to PLSR and PCR, whose $\mathbb{R}^2$ values are very close to 1. Nonetheless, the computational times of training and testing datasets for LSSVR are much longer than those of PLSR and PCR. Consequently, a highly accurate LSSVR model-based YI sensor shows promising applications particularly if the computational load can be further minimized.

Keywords: Least square support vector regression, Yellow color, Model, Algorithm, Yellowness Index

1. INTRODUCTION

Current modern manufacturing processes have a strong requirement for effective and rapid computing modeling, such as soft sensors to monitor, control, optimize, and automate their chemical processes to enhance Industry 4.0. For instance, the purities of fatty acid and glycerol via hydrolysis of palm oil are required to be monitored and optimized to have a better quality [1], [2]. Hence, a data strategy that systematically connects industry 4.0 technologies and data analysis with sector-specific competencies, especially since it is associated with the newest advances in soft sensors. Moreover, a soft sensor is one of the advances in modern statistics and computing technology that can bring new chances to form novel or improved evaluation techniques [3]. A soft sensor utilizes regression analysis that involves a set of machine learning techniques allowing the prediction of a continuous output variable based on the value of one or multiple predictor variables or input variables [4], [5]. To be more specific, the soft sensor model is used for the forecast of future targeted data based on the correlation between individual variables to have an optimized and productive operating process when it is incorporated into a processing system [6]. To date, soft sensor models have been widely studied to predict accuracy for the targeted and desired variable for superior chemical process monitoring and controlling a process to produce a good quality product [7].

The least-square support vector regression (LSSVR) was introduced by Suykens, et al. [8] is one of the regression models that is modified from support vector regression (SVR). The difference between LSSVR from SVR is that it solves a set of linear systems to obtain the solution instead of solving a quadratic program problem. Hence, LSSVR is simpler and has a lower computational load as compared to SVR. The effectiveness of LSSVR has been widely studied by Yu, et al. [9] in measuring the ecological parameters of rice wine, Sivaramakrishnan, et al. [10] in estimating the concentrations of a product in an acid-catalyzed propylene oligomerization, and Yeo and Lau [11] in forecasting the whiteness index of fabric. Meanwhile, these studies also included the predictive performance comparison of LSSVR with partial least square regression (PLSR) which is another frequently used regression model. However, they did not consider the prediction of the yellowness index (YI) using LSSVR and the comparison of LSSVR and PLSR on YI prediction has not been carried out.

On the other hand, both PLSR and principal component regression (PCR) become popular models in estimating the quality, and process monitoring in a process control system due to their simplicities [12], [13]. Moreover, these two

E-mail address: christineyeo@curtin.edu.my, agus.saptoro@curtin.edu.my

models have been studied by many researchers such as Singh and Sarkar [14], Ergon, et al. [15], Ergon [16], and Engelen, et al. [17]. PCR is an extension regression analysis method of PCR that consists of principal component analysis (PCA). This PCA model is a famous unsupervised dimensionality reduction mathematical technique that separates data where perceptions can be represented by the collective relation of the input variable that connects to the current PLSR application [18]. On the other hand, PCR searches a small scale space grabbing the optimum estimation of the dissimilarity in independent variables, X. Meanwhile, the PLSR technique is the same as PCR but PLSR differs by including the measurement of the dependent and independent variables to predict the targeted output variable [12].

Nevertheless, PLSR discovers hybrid features from PCA and multiple linear regression [19]. PLSR-based models are usually derived using Nonlinear Iterative Partial Least Squares (NIPALS) and singular value decomposition (SVD). NIPALS performs superior to SVD since it is faster, can handle missing data, and determines the components sequentially [20], [21]. Thus, the majority of PLSR-based algorithms are developed based on the NIPALS-based algorithm [22]. Since the invention of NIPALS, PLSR has been recognized as a popular model for modeling and analyzing huge multivariable collinear sets of data [23] since it is a dimension-reduction method. Additionally, PLSR has been applied for data analysis in various fields such as the dynamic monitoring oxidation process of nut oils done by Wang, et al. [24], hyperspectral prediction of soil organic matter content studied by Shen, et al. [25], and forecasting chemical properties of fish muscle conducted by Cheng and Sun [26]. However, a limited color-related study is using the abovementioned PLSR, PCR, and LSSVR. The color index like YI is a reference for all colorants such as dye, and pigment [27] [27] and it is a simple numerical expression that determines the color of an object. YI is a value that denotes the departure level of a sample's color from a desired white and becomes yellow [28], [29].

Due to the importance of color in a range of products and technologies, color engineering has been established to synthesize the disciplines of color science and computer engineering [30]. Moreover, color engineering is essential in the development of color imaging products in some big companies like Google, Apple, Samsung, Amazon, etc. Besides, color is a critical quality attribute in industries like paints, chemicals, and agriculture industries ensuring consistent color in products is crucial for meeting customer expectations and maintaining product quality. This is because the color is an organoleptic characteristic that affects consumers' preference to adopt an item. Furthermore, color has a strong connection to product quality. Hence, color has been studied by many researchers [31], [32], [33], [34]. Despite that, the majority of studies used expensive color measurement instruments including colorimeter, and spectrophotometer. Moreover, they can only determine the color space data for color index calculation, and these instruments cannot measure the color index directly. On top of that, YI was studied by Blanco, et al. [35] and Digesù, et al. [36] to investigate the wheat quality. However, research studies focusing on YI, especially using soft sensor models, remain limited. In other words, YI is computed from color scale parameters that assess perceptual yellowness. However, the color data to calculate YI values is usually measured by expensive hardware instruments. Furthermore, a direct and rapid YI estimation utilizing this hardware equipment is impossible. And, soft sensors models could be a potential solution to this problem statement.

To address the abovementioned research gaps, the current study focuses on obtaining a low-cost, rapid, and straightforward YI measurement technique. This comparative study was carried out to examine the performances of the PLSR, LSSVR, and PCR methods to forecast the YI using the color space data. Then, an RGB-listed standard color chart was employed to collect the dataset for the LSSVR, PLSR, and PCR models. Later, the RGB can be converted into CIE LAB and CIE XYZ [37]. The YI values can be obtained from the CIE XYZ. The performances of LSSVR, PLSR, and PCR models were measured and distinguished. In other words, the outcomes of the performances for LSSVR, PLSR, and PCR including coefficients of determination ($\mathbb{R}^2$), root mean square error (RMSE), mean absolute errors (MAE), and computation times (t) were adopted and compared among them.

2. MATERIALS AND METHODS

This section is divided into seven sub-sections to explain the research materials and methods that were used in this study. Firstly, this section describes the LSSVR method. Next, it is followed by data splitting, parameter setting, different color space conversions, as well as YI computation. Then, data generation, predictive performance analysis, and computer configuration are described.

A. Least square support vector regression model

As compared to the SVR that has quadratic programming, LSSVR provides a more effective solution with lower computational complexity and time since LSSVR solves a set of linear equations [38]. The LSSVR with radial basis function (RBF) kernel is utilized in this study. Kernel functions have been utilized for pattern recognition, regression, approximation, and operator inversion [39]. The effectiveness of LSSVR has been recently proven by Pervez, et al. [40], Ngu and Yeo [41], and Yeo and Lau [11]. This LSSVR was introduced by Xu, et al. [42] and it adopts the multi-output case that was conducted by An, et al. [43]. And this multi-output setting in this LSSVR makes it provides a better training algorithm. The LSSVR model is briefly explained in this section and it was taken from Xu, et al. [42] and Yeo and Lau [11]. Let $Y = [y_{i,j}] \in \Re^{l \times m}$, $y_{i,j}$ is the (i,j)-th of output with an l×m matrix. For a provided total number of datasets, N_T, i.e., $\{(x_i, y^i\}_{i=1}^{l}$ in which $x_i \in \Re^d$ is the input vector and $y^i \in \Re^m$ is the output vector. Moreover, the objective of the LSSVR is to estimate an output vector $y \in \Re^m$ from a provided input vector $x \in \Re^d$ in which this regression issue is able to form by learning a mapping from $\Re^d$ to $\Re^m$. Then, this LSSVR provides solutions for the regression issue by finding the weighted value vector, $W = (w_1, w_2, ..., w_m) \in \Re^{n_h \times m}$ and a threshold value, $b = (b_1, b_2, ..., b_m)^T \in \Re^m$ to meet the following targeted objective function with its constraints in 1 and 2:

$$\min_{W \in \Re^{m \times n_h}, b \in \Re^m \frac{1}{2}^T \frac{1}{2}^T} \tag{1}$$

$$s.t. Y = Z^T W + repmat(b^T, l, 1)] + \Xi \tag{2}$$

where γ is a non-negative real regularised parameter, $\xi = (\xi_1, \xi_2, ..., \xi_l)^T \in \Re^l$ is a vector having slack variables, $Z = (\phi(x_1), \phi(x_2), ..., \phi(x_l)) \in \Re^{n_h \times l}, \phi : \Re^d \rightarrow \Re^{n_h}$ is a data mapping to an infinite dimensional Hilbert feature space through the nonlinear mapping function ϕ with n_h dimensions, and $\Xi = (\xi_1, \xi_2, ..., \xi_m) \in \Re_+^{l \times m}$ is an $l \times m$ order of matrix containing slack variables with $\Re_+$. More details about this LSSVR model can be found in Xu, et al. [42] [42].

The RBF that is employed as a kernel function in the LSSVR is taken from Keerthi and Lin [44]. It is displayed in 3 and is integrated into the LSSVR.

$$k(x, z) = e^{(-p\|x-z\|^2)} \tag{3}$$

where p is the non-negative hyperparameter of the RBF kernel function. Furthermore, LSSVR has three tuning parameters involving γ, λ, and p that are tuned via the leave-one-out (LOO) technique to achieve the mean relative error, δ that is displayed in 4.

$$\delta = \frac{1}{l} \sum_{i=1}^{l} \frac{|Y_i - \hat{Y}_i|}{Y_i} \tag{4}$$

where Y_i and $\hat{Y}_i$ are the real and the predicted outputs, respectively.

B. Splitting of data and parameters setting

In this study, there were 76 datasets consisting of the values of YI, XYZ, RGB, and LAB were gathered from the RGB-listed color charts. Later, the collected datasets were transferred into MATLAB software, and then 75% of the training and 25% of the testing data were utilized [45]. The training data was utilized to formulate LSSVR, PLSR, and PCR models. Later, these split datasets were run using these three regression models, and then model evaluations were carried out [46]. Then, the values of RMSE, MAE, $\mathbb{R}^2$, and computational times, t in seconds for all models were obtained. All regression model values adopted in the three are shown in Table I. The total number of the dataset, N_T was split into a number of the training dataset, N_1, and a number of the testing dataset, N_2. Moreover, LV which is the number of LV was set as 1. Meanwhile, γ, λ, and p in the LSSVR model were tuned using the LOO method. The estimation of the LOO method for these three feature values is calculated from the training datasets and they are fixed for both training and testing datasets.

C. Color spaces transformations and yellowness index computation

RGB is the opposite of cyan, magenta, yellow, and key (black) (CMYK) which is another color space because it is an "additive" process. RGB utilizes white as an integration of all primary colors and black without the presence of light. On the contrary, CMYK employs white as the natural color of the print background and black as the integration of colored inks. RGB color space is recognized as the famous color space and RGB is integrated to form a color [47]. Nonidentical color needs not to have the same quantity of RGB, and the amount required for R, G, and B to build color is denoted as tristimulus values that are X, Y, and Z [48]. They are intermediate color spaces and the coordinating of RGB values that demonstrate any available color with the lightness information involved [49]. Barbero-Álvarez, et al. [49] reported that the XYZ values can be calculated from the RGB values using 5 to 7.

$$X = 0.4125R + 0.3576G + 0.1804B \tag{5}$$

$$Y = 0.2127R + 0.7152G + 0.0722B \tag{6}$$

$$Z = 0.0193R + 0.1192G + 0.9503B \tag{7}$$

Particularly, YI can be obtained from the American Society for Testing and Materials 313-00 which is described as 8 [34].

$$YI = \frac{(100(1.3013X - 1.1498Z))}{Y} \tag{8}$$

Moreover, XYZ values are used to estimate the CIE L * a * b * color values [50]. 9 and 10 explain these CIE L * a * b * color values that are normalized to the reference white [51].

$$L* = 116\sqrt[3]{\frac{Y}{Y_n}} - 16$$
$$a* = 500(\sqrt[3]{\frac{X}{X_n}} - \sqrt[3]{\frac{Y}{Y_n}})$$
$$b* = 200(\sqrt[3]{\frac{Y}{Y_n}} - \sqrt[3]{\frac{Z}{Z_n}})$$

(9)

where the $\frac{X}{X_n}$, $\frac{Y}{Y_n}$, and $\frac{Z}{Z_n}$ are larger than 0.008856.

$$L* = 903.3(\frac{Y}{Y_n})$$
$$a* = 500\{[7.87(\frac{X}{X_n}) + \frac{16}{116}] - [7.87(\frac{Y}{Y_n}) + \frac{16}{116}]\}$$
$$b* = 200\{7.87(\frac{X}{X_n}) + \frac{16}{116}] - [7.87(\frac{Z}{Z_n}) + \frac{16}{116}]\}$$

(10)

TABLE I. The parametric quantity and amounts employed for the least-square regression methods

Parametric quantity	N_T	N_1	N_2	LV	γ	λ	p
Amounts	76	57	19	1	15	10	3

where the $\frac{X}{X_n}$, $\frac{Y}{Y_n}$, and $\frac{Z}{Z_n}$ are equal to or lesser than 0.008856.

D. Data generation

An RGB-listed standard color chart was utilized to gather data for formulating and testing the used methods in this study. The RGB values on the color chart were used to calculate the YI values. These RGB values were initially input into Eqs. (12) to (14). Next, the obtained XYZ values were filled into eq. (15) to calculate the YI. Then, with these XYZ values, Eqs. (16) and (17) are used to get the CIE LAB values. In LSSVR, PLSR, and PCR models, the collected RGB, calculated XYZ and obtained CIE LAB values are set as the input variables. And then, YI values are denoted as the output variable. Table II shows the sample of input data from datasets used for the regression models.

E. The predictive performance analysis and computer configurations

As mentioned earlier, the outcomes of the predictive performances for LSSVR, PLSR, and PCR are $\mathbb{R}^2$, RMSE, MAE, and t. RMSE is a conventional method used in a model evaluation [21], [52] that obtains the differences between the real and predicted values using 11. When RMSE values are lower, they indicate a better predictive performance of the model [6], [53], [54], [55]. Aside from that, $\mathbb{R}^2$ as shown in 12 is also used [56], [57]. The closer the $\mathbb{R}^2$ value is to 1, the better the fit.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_T}(y_i - \hat{y_i})^2}{N_T}} \tag{11}$$

where Y_i and $\hat{Y}_i$ are the real and estimated output variables, respectively.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(\hat{y}_i - \bar{y_i})^2}$$

(12)

where $\bar{y}$ is the mean value of y_i.

Moreover, in this study, MAE is also used to investigate the prediction of the regression models. MAE measures the mean magnitude of errors in a set of predictions regardless of the direction. The equation of MAE can be seen in 13 [58].

$$MAE = \sum \frac{|y_i - \hat{y_i}|}{N_T}$$

(13)

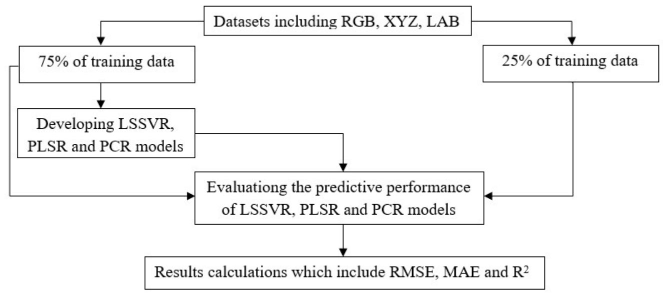

Figure 1. A flow diagram illustrating the data pipeline from the datasets to the obtained results.

Besides, computational times of all regressions for both training and a testing dataset which are denoted as t_1 and t_2, respectively, are analyzed and compared. On the other hand, the LSSVR, PLSR, and PCR models were formulated and run on a laptop, namely Asus ZenBook UX305 and MATLAB version 2021. This laptop is equipped with 64-bit Windows 10; a 2.20 GHz Intel Core M3-6Y30 central processing unit processor, 4.0 GB of Random Access Memory, and 128 GB solid-state drive storage. Figure 1 displays a flow diagram illustrating the data pipeline from the datasets to the obtained results.

3. RESULTS AND DISCUSSIONS

As mentioned earlier, the LSSVR, PLSR, and PCR models are utilized to predict the YI values for the collected datasets. Then, the outcomes from these regression algorithms are depicted in Table III. From Table III, $RMSE_1$, $\mathbb{R}^2$, and MAE_1 denote the RMSE, $\mathbb{R}^2$, and MAE for training data. Meanwhile, $RMSE_2$, $\mathbb{R}^2$, and MAE_2 are referring to the RMSE, $\mathbb{R}^2$, and MAE for testing data. Furthermore, t_1 is the computational time for training data while t_2 is the computational load for testing data. From Table III, notice that LSSVR gives the lowest RMSE and MAE values, and it also has the highest $\mathbb{R}^2$ and t values. To be more specific, the $RMSE_1$ for LSSVR is 261,406% and 294,218% lower than PLSR and PCR while its $RMSE_2$ values are 725% and 772% smaller than PLSR and PCR. Also, it's MAE_1 is lowered by 253,946% and 287,554% as compared to PLSR and PCR. Meanwhile, the MAE_2 for LSSVR is 1,365% which is 1,512% smaller than PLSR and PCR. Furthermore, LSSVR has 3% to 8% higher $\mathbb{R}^2$ for training and testing data than PLSR and PCR. This is due to the RBF kernel function in the LSSVR which can coordinate the data into higher dimensional space to have a superior predicted performance [11], [59], [60]. Meanwhile, LSSVR uses the LOO technique to tune the parameters to get optimal results. However, due to the presence of RBF kernel function and

TABLE II. A table showing the sample of input data from datasets

R	G	B	X	Y	Z	L	A	B	YI
255	254	238	2.51	2.53	2.40	99.35	-1.96	6.82	25.9
255	252	219	2.47	2.50	2.23	98.43	-3.97	15.29	25.9
255	251	200	2.43	2.48	2.07	97.80	-6.22	24.15	31.8
255	247	157	2.33	2.42	1.69	96.09	-9.50	43.48	45.4
255	245	134	2.28	2.39	1.48	95.28	-10.76	53.57	52.9

LOO technique, the computational times which are t_1 and t_2 for LSSVR are 100% higher than PLSR and PCR models.

From Table III, the PLSR has $RMSE_1$ and $RMSE_2$ which are 12.55% and 5.68% smaller than PCR while the MAE_1 and MAE_2 for PLSR are also 13.23% and 10.03% lower than PCR. Moreover, the $\mathbb{R}^2$ values for PLSR are approximately 1% higher than PCR. The main reason for this result is due to the PLSR transformation objective being the optimum intersection between the output and input variables, however, PCR has optimized the block coverage within the input variables only [61]. Furthermore, PLSR also finds for the LVs that compute a simultaneous decomposition of input and output variables. These LVs explain the maximum covariance between the input and output variables [19]. Besides, both PLSR and PCR models have low computational burdens where their t_1 and t_2 are less than 1 sec. Additionally, Figures 2(i) and 2(ii) show the comparison of the real YI and the estimated YI values from LSSVR, PLSR, and PCR models for both datasets.

Notice that the predicted output which is the YI from the LSSVR is perfectly fixing the true data, hence it covers the true data which is a blue line in Figure 2(i). Moreover, it can be seen in Figures 2(i) and 2(ii) that all YI values are non-negative. In reality, the YI value for a yellow color sample must be non-negative and the non-positive YI values that do not represent yellow color are for a bluish object [62]. Hence, these non-negative YI values are displayed in Figures 3(i) and 3(ii) and also validated that the RGB values on the standard color chart are for yellow color objects and that the produced YI values are reasonable. Notice that these true and predicted YI values from all used regression models are closed while the $\mathbb{R}^2$ values for all regression models shown in Table III are more than 0.9. The high values of $\mathbb{R}^2$ for LSSVR, PLSR, and PCR indicate that these regression models fix the YI data well and can be employed to estimate the YI values [63]. Despite that, the LSSVR has the highest values of $\mathbb{R}^2$ for both datasets as displayed in Table III as well as Figures 3(i) and 3(ii). Hence, in conclusion, the LSSVR has better predictive performance on YI values than the PLSR and PCR models.

4. Conclusions and Future Work

In this study, a comparative study was performed to examine the LSSVR, PLSR, and PCR models' performance in the estimation of the YI values. Moreover, the dataset was collected using a standard RGB-listed color chart.

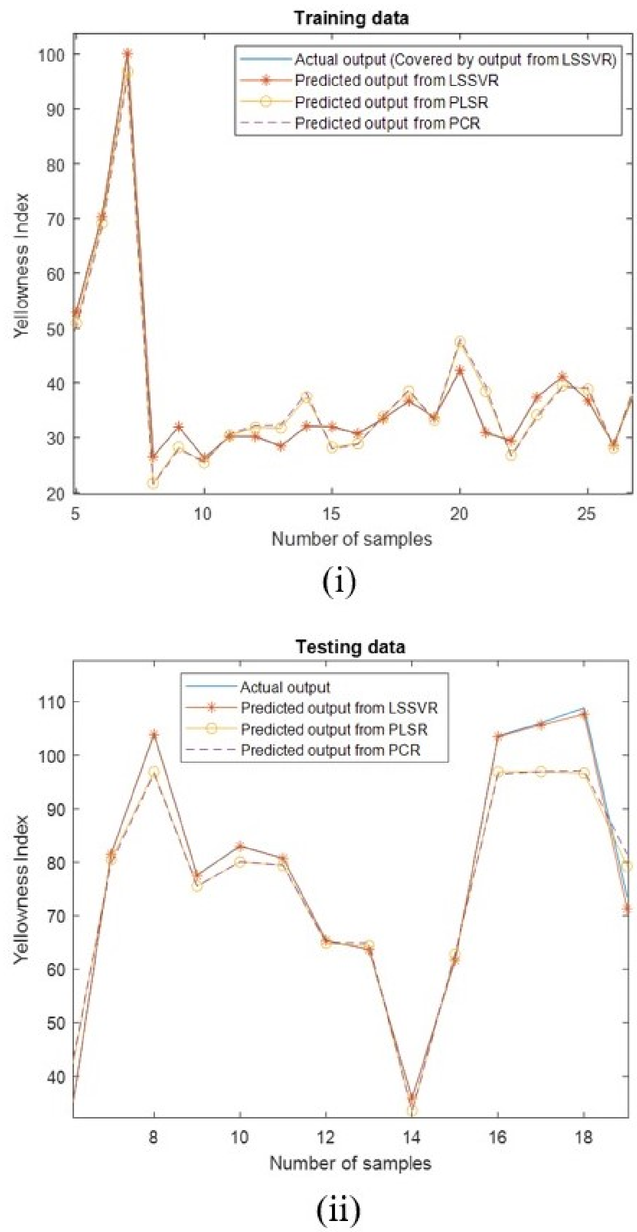

Figure 2. The comparison of the true values and the estimated YI values from LSSVR, PLSR, and PCR models for, (i) Training data, and (ii) Testing data.

TABLE III. The obtained outcomes were from PLSR, LSSVR, and PCR methods

Results	LSSVR	PLSR	PCR
$RMSE_1$	0.0015	3.8159	4.2947
MAE_1	0.0012	2.9381	3.3268
R_1^2	1.0000	0.9680	0.9592
t_1 (sec)	10.9428	0.0052	0.0039
$RMSE_2$	0.5962	4.9203	5.2000
MAE_2	0.2360	3.4573	3.8040
R_2^2	0.9992	0.9269	0.9181
t_2 (sec)	10.8914	0.0032	0.0032

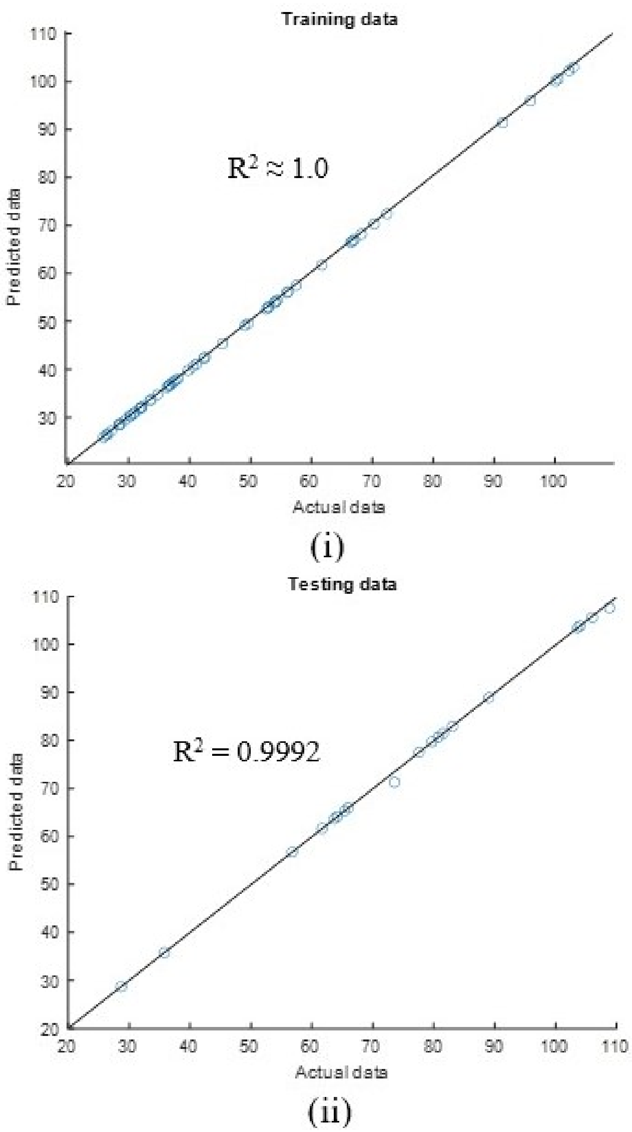

Figure 3. Correlation between the true and estimated YI values for, (i) Training data, and (ii) Testing data.

This collected dataset was used as both datasets for the PLSR, LSSVR, and PCR models development and evaluation. Then, the predictive outcomes from PLSR, LSSVR, and PCR methods on YI estimation were evaluated and compared. From the results, notice that LSSVR performed much superior to PLSR and PCR on the forecasting of YI values. The outcomes show that the RMSE values of both datasets for LSSVR are roughly 725% to 294,218% smaller than PLSR and PCR. Moreover, the MAE values of both datasets for LSSVR are 1,365% to 287,554% smaller than PLSR and PCR. The computational times for LSSVR, PLSR, and PCR models are relatively low, which are less than 11 sec.

Nevertheless, the computational times for both datasets for LSSVR are 100% much higher than both PLSR and PCR. Despite that, the $\mathbb{R}^2$ values of both datasets for LSSVR are higher and closer to 1 as compared to PLSR and PCR. Moreover, $\mathbb{R}^2$ values for PLSR, LSSVR, and PCR are greater than 0.9 which denotes lesser dissimilarities between the true and predicted YI values from these regression models. Hence, these regression models can be utilized to estimate the YI on a yellowish sample and be an alternative to the pricy color measurement instruments such as spectrophotometers and colorimeters.

YI is computed from color scale parameters that assess perceptual yellowness [64]. However, the color data to calculate YI values is usually measured by expensive hardware instruments. Furthermore, a direct and rapid YI estimation utilizing this hardware equipment is impossible. The purchasing cost of a spectrophotometer is roughly \$5,000 while a colorimeter costs between \$5,500 to \$6,500 [65], [66]. From the results, LSSVR performed much better than PLSR and PCR, and hence it is more suitable to be integrated with the MATLAB mobile apps. Hence, this research study can introduce a novel tool to measure the YI value of an object. On the other hand, the future study can investigate the performance of the LSSVR model in real-time in YI values measurement using the existing MATLAB mobile app where the different light intensity levels, denoising of the taken images, image enhancement, and image restoration can be further investigated.

ACKNOWLEDGMENT

The authors thank Curtin University Malaysia for giving financial support for this study. This research is funded by the Early Career Researchers Startup Fund 2021 (Grant number: 100040) from the Research and Development Office at Curtin University Malaysia. The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS STATEMENT

Wan Sieng Yeo collected samples, validated methodology, conducted simulations, did data analysis, and wrote, reviewed, and edited the paper. Agus Saptoro helped to edit and improve the paper.

ORCID

Wan Sieng Yeo® https://orcid.org/0000-0003-3248-3521

Agus Saptoro® https://orcid.org/0000-0002-1734-4788

References

[1] S. C. N. F. S. A. K. M. S. A. M. M. C. Y. L. Toon Ming Foo, Tien Pei Kim and W. S. Yeo, "A new conceptual process design and economic analysis of a fatty acids and glycerine production plant using palm oil," *Chemical Papers*, vol. 76, pp. 3471–3483, 2022. [Online]. Available: Available:https://doi.org/10.1007/s11696-022-02102-6

[2] A. Y. Wan Sieng Yeo, "The external mass transfer model for the hydrolysis of palm olein using immobilized lipase in a scaled-up recirculated packed-bed reactor," *Journal of Environmental Chemical Engineering*, vol. 7, no. 4, p. 103185, 2019. [Online]. Available: Available:https://doi.org/10.1016/j.jece.2019.103185

[3] Y. F. M. L. Ming-feng Zhan, Zong-wu Cai, "Recent advances in statistical methodologies in evaluating program for high-dimensional data," *Applied Mathematics-A Journal of Chinese Universities*, vol. 37, no. 1, pp. 131–146, 2022. [Online]. Available: Available:https://doi.org/10.1007/s11766-022-4489-3

[4] P. K. M. K. Wan Sieng Yeo, Agus Saptoro, "Just-in-time based soft sensors for process industries: A status report and recommendations," *Journal of Process Control*, vol. 128, p. 103025, 2023. [Online]. Available: Available:https://doi.org/10.1016/j.jprocont.2023.103025

[5] Z. Y. C. J. N. A. S. Jameson Malang, Wan Sieng Yeo, "A comparison study between different kernel functions in the least square support vector regression model for penicillin fermentation process," *MATEC Web of Conferences*, vol. 377, p. 01025, 2023. [Online]. Available: Available:https://doi.org/10.1051/matecconf/202337701025

[6] W. S. Yeo, "Prediction of yellowness index using partial least square regression model," *International Conference on Green Energy, Computing and Sustainable Technology (GECOST), IEEE Xplore*, pp. 1–5, 2020. [Online]. Available: Available:https://doi.org/10.1109/GECOST52368.2021.9538723

[7] W. S. Y. Joyce Chen Yen Ngu, "A comparative study of different kernel functions applied to lw-kpls model for nonlinear processes," *Biointerface Research in Applied Chemistry*, vol. 13, no. 2, pp. 1–16, 2023. [Online]. Available: Available:https://doi.org/10.33263/BRIAC132.184

[8] J. D. B. B. D. M. J. P. V. Johan AK Suykens, Tony Van Gestel, "Least squares support vector machines," *World Scientific*, pp. 1–308, 2002. [Online]. Available: Available:https://doi.org/10.1142/5089

[9] H. X. Y. Y. B. L. X. P. Haiyan Yu, Hongjian Lin, "Prediction of enological parameters and discrimination of rice wine age using least-squares support vector machines and near infrared spectroscopy," *Journal of Agricultural Food Chemistry*, vol. 56, no. 2, pp. 307–313, 2008. [Online]. Available: Available:https://doi.org/10.1021/jf0725575

[10] A. d. K. V. P. Kaushik Sivaramakrishnan, Jinjun Nie, "Least squares-support vector regression for determining product concentrations in acid-catalyzed propylene oligomerization," *Industrial Engineering Chemistry Research*, vol. 57, no. 39, pp. 13 156–13 176, 2018. [Online]. Available: Available:https://doi.org/10.1021/acs.iecr.8b02539

[11] W. J. L. Wan Sieng Yeo, "Predicting the whiteness index of cotton fabric with a least squares model," *Cellulose*, vol. 28, pp. 8841–8854, 2021. [Online]. Available: Available:https://doi.org/10.1007/s10570-021-04096-y

[12] Y. H. G. C. L Lu Zheng, Thomas J Mc Avoy, "Application of multivariate statistical analysis in batch processes," *Industrial Engineering Chemistry Research*, vol. 40, no. 7, pp. 1641–1649, 2001. [Online]. Available: Available:https://doi.org/10.1021/ie000630+

[13] Z. S. Junhua Zheng, "Linear subspace principal component regression model for quality estimation of nonlinear and multimode industrial processes," *Industrial Engineering Chemistry Research*, vol. 56, no. 21, pp. 6275–6285, 2017. [Online]. Available: Available:https://doi.org/10.1021/acs.iecr.7b00498

[14] A. S. Manjeet Singh, "Comparative study of the plsr and pcr methods in laser-induced breakdown spectroscopic analysis," *Journal of Applied Spectroscopy*, vol. 85, no. 5, pp. 962–970, 2018. [Online]. Available: Available:https://doi.org/10.1007/s10812-018-0746-x

[15] G. A. Rolf Ergon, D Granato, "Principal component regression (pcr) and partial least squares regression (plsr)," *in Mathematical Statistical Methods in Food Science Technology G. A. Daniel Granato*, pp. 121–142, 2014. [Online]. Available: Available:https://doi.org/10.1002/9781118434635

[16] R. Ergon, "Finding y-relevant part of x by use of pcr and plsr model reduction methods," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 21, no. 12, pp. 537–546, 2007. [Online]. Available: Available:https://doi.org/10.1002/cem.1062

[17] K. V. B. S. V. S Engelen, Mia Hubert, "Robust pcr and robust plsr: a comparative study," *Theory and Applications of Recent Robust Methods. Statistics for Industry and Technology*, vol. 8, pp. 105–117, 2004. [Online]. Available: Available:https://doi.org/10.1007/978-3-0348-7958-3_10

[18] L. W. Hervé Abdi, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010. [Online]. Available: Available:https://doi.org/10.1002/wics.101

[19] H. Abdi, "Partial least square regression (pls regression)," *Encyclopedia for Research Methods for the Social Sciences*, vol. 6, no. 4, pp. 792–795, 2003. [Online]. Available: Available:http://www.utd.edu/herve

[20] A. P. F.-J. G.-D. Sandra Ramírez, Manuel Zarzo, "A methodology for discriminant time series analysis applied to microclimate monitoring of fresco paintings," *Sensors*, vol. 21, no. 2, p. 436, 2021. [Online]. Available: Available:https://doi.org/10.3390/s21020436

[21] A. S. Vivianna Maria Mickel, Wan Sieng Yeo, "Evaluating the performance of newly integrated model in nonlinear chemical process against missing measurements," *Chemical Product and Process Modeling*, vol. 14, no. 4, pp. 1–20, 2019. [Online]. Available: Available:https://doi.org/10.1515/cppm-2018-0066

[22] B. A. Tahir Mehmood, "The diversity in the applications of partial least squares: an overview," *Journal of Chemometrics*, vol. 30, no. 1, pp. 4–17, 2016. [Online]. Available: Available:https://doi.org/10.1002/cem.2762

[23] J. F. M. Johan A Westerhuis, Theodora Kourti, "Analysis of multiblock and hierarchical pca and pls models," *Journal of chemometrics: A Journal of the chemometrics society*, vol. 12, no. 5, pp. 301–321, 1998. [Online]. Available: Available:https://doi.org/10.1002/(SICI)1099-128X(199809/10)12:5⟨301::AID-CEM515⟩3.0.CO;2-S

[24] Y. Z. Y. C. W. Y. H. Y. Y. G. Y. X. Cheng Wang, Yingying Sun, "Dynamic monitoring oxidation process of nut oils through raman technology combined with plsr and rf-plsr model," *LWT*, vol. 146, p. 111290, 2021. [Online]. Available: Available:https://doi.org/10.1016/j.lwt.2021.111290

[25] J. Y. Z.-L. L. P. L. Q. Y. S.-B. D. Lanzhi Shen, Maofang Gao, "Hyperspectral estimation of soil organic matter content using different spectral preprocessing techniques and plsr method," *Remote Sensing*, vol. 12, no. 7, p. 1206, 2020. [Online]. Available: Available:https://doi.org/10.3390/rs12071206

[26] D.-W. S. Jun-Hu Cheng, "Partial least squares regression (plsr) applied to nir and hsi spectral data modeling to predict chemical properties of fish muscle," *Food Engineering Reviews*, vol. 9, no. 1, pp. 36–49, 2017. [Online]. Available: Available:https://doi.org/10.1007/s12393-016-9147-1

[27] E. Wich, "The colour index," *Color Research and Application*, pp. 77–80, 1977. [Online]. Available: Available:https://doi.org/10.1002/col.5080020205

[28] R. D. Paravina, "New shade guide for tooth whitening monitoring: visual assessment," *The Journal of Prosthetic Dentistry*, vol. 99, no. 3, pp. 178–184, 2008. [Online]. Available: Available:https://doi.org/10.1016/S0022-3913(08)60041-4

[29] A. N. S. Wan Sieng Yeo, "A yellowness index simulation model," *IEEE Xplore*, pp. 1–4, 2021. [Online]. Available: Available:https://doi.org/10.1109/GECOST52368.2021.9538636

[30] L. M. Phil Green, "Colour engineering: achieving device independent colour," *Journal of the Society for Information Display*, vol. 13, no. 3, pp. 663–665, 2011. [Online]. Available: Available:https://doi.org/10.1889/1.1824248

[31] C. D.-R. M. G. L.-R. E. M. R.-R. Miguel Padrón-Mederos, Beatriz Rodríguez-Galdón, "Quality evaluation of minimally fresh-cut processed pineapples," *LWT*, vol. 129, p. 109607, 2020. [Online]. Available: Available:https://doi.org/10.1016/j.lwt.2020.109607

[32] M. I. P. A. V. Tatiana Dumitra Panaite, Silvia Mironeasa, "Liquid egg products characterization during storage as a response of novel phyto-additives added in hens diet," *Emirates Journal of Food Agriculture*, vol. 31, no. 4, pp. 304–314, 2019. [Online]. Available: Available:http://doi.org/10.9755/ejfa.2019.v31.i4.1937

[33] S. Y. Huseyin Goger, Sahnur Erdogan Demirtas, "A selection study for improving eggshell colour in two parent lines of laying hens and their hybrids," *Italian Journal of Animal Science*, vol. 15, no. 3, pp. 390–395, 2016. [Online]. Available: Available:https://doi.org/10.1080/1828051X.2016.1215232

[34] L. Z. C.-Z. J. Y. F. S. W. Qingguo Wang, Yu Cao, "Effects of postharvest curing treatment on flesh colour and phenolic metabolism in fresh-cut potato products," *Food Chemistry*, vol. 169, pp. 246–254, 2015. [Online]. Available: Available:https://doi.org/10.1016/j.foodchem.2014.08.011

[35] P. C. A. G. G. M. A. S. R. S. A. M. D. P. D. V. A. M. M. Blanco, Antonio and L. Cattivelli, "Quantitative trait loci for yellow pigment concentration and individual carotenoid compounds in durum wheat," *Journal of Cereal Science*, vol. 54, no. 2, pp. 255–264, 2011. [Online]. Available: Available:https://doi.org/10.1016/j.jcs.2011.07.002

[36] L. C. G. M. A. B. AM Digesù, C Platani, "Genetic variability in yellow pigment components in cultivated and wild tetraploid wheats," *Journal of Cereal Science*, vol. 50, no. 2, pp. 210–218, 2009. [Online]. Available: Available:https://doi.org/10.1016/j.jcs.2009.05.002

[37] J. A. S. C. S. S. C. V. Manuel Jesús Sánchez Chero, William Rolando Miranda Zamora, "Application of the computer vision system to the measurement of the cie l* a* b* color parameters of fruits," *in International Conference on Applied Human Factors and Ergonomics*, vol. 1213, pp. 341–347, 2020. [Online]. Available: Available:https://doi.org/10.1007/978-3-030-51328-3_47

[38] C. Y. Hongbin Liu, "A robust localized soft sensor for particulate matter modeling in seoul metro systems," *Journal of Hazardous Materials*, vol. 305, pp. 209–218, 2016. [Online]. Available: Available:https://doi.org/10.1016/j.jhazmat.2015.11.051

[39] B. S. Alexander J Smola, "On a kernel-based method for pattern recognition, regression, approximation, and operator inversion," *Algorithmica*, vol. 22, pp. 211–231, 1998. [Online]. Available: Available:https://doi.org/10.1007/PL00013831

[40] F. S. M. M. J. Z. S. M. R. L. L. X. X. V. N. Y. C. Md Nahid Pervez, Wan Sieng Yeo, "Sustainable fashion: Design of the experiment assisted machine learning for the environmental-friendly resin finishing of cotton fabric," *Heliyon*, vol. 9, no. 1, p. e12883, 2023. [Online]. Available: Available:https://doi.org/10.1016/j.heliyon.2023.e12883

[41] W. S. Y. Joyce Chen Yen Ngu, "Prediction of dissolved oxygen using least square support vector regression model," *IEEE Xplore*, pp. 70–74, 2023. [Online]. Available: Available:https://doi.org/10.1109/GECOST55694.2022.10010638

[42] X. Q. L. Z. L. L. Shuo Xu, Xin An, "Multi-output least-squares support vector regression machines," *Pattern Recognition Letters*, vol. 34, no. 9, pp. 1078–1084, 2013. [Online]. Available: Available:https://doi.org/10.1016/j.patrec.2013.01.015

[43] L.-D. Z. S.-G. S. Xin An, Shuo Xu, "Multiple dependent variables ls-svm regression algorithm and its application in nir spectral quantitative analysis," *Spectroscopy and Spectral Analysis*, vol. 29, no. 1, pp. 127–130, 2009. [Online]. Available: Available:https://doi.org/10.3964/j.issn.1000-0593(2009)01-0127-04

[44] C.-J. L. S Sathiya Keerthi, "Asymptotic behaviors of support vector machines with gaussian kernel," *Neural Computation*, vol. 15, no. 7, pp. 1667–1689, 2003. [Online]. Available: Available:https://doi.org/10.1162/089976603321891855

[45] S. S. S. R. S. S. K. Sudhakar Singha, Srinivas Pasupuleti, "Prediction of groundwater quality using efficient machine learning technique," *Chemosphere*, vol. 276, p. 130265, 2021. [Online]. Available: Available:https://doi.org/10.1016/j.chemosphere.2021.130265

[46] J. Z. Jiandong Huang, Yuantian Sun, "Reduction of computational error by optimizing svr kernel coefficients to simulate concrete compressive strength through the use of a human learning optimization algorithm," *Engineering with Computers*, vol. 38, pp. 3151–3168, 2021. [Online]. Available: Available:https://doi.org/10.1007/s00366-021-01305-x

[47] F. A. K. Dibya Jyoti Bora, Anil Kumar Gupta, "Comparing the performance of l* a* b* and hsv color spaces with respect to color image segmentation," *International Journal of Emerging Technology and Advanced Engineering*, vol. 5, no. 2, pp. 1–12, 2015. [Online]. Available: Available:https://doi.org/10.48550/arXiv.1506.01472

[48] N. L. W. W. Y. L. Q. F. Chenyang Deng, Ningfang Liao, "Reconstruction of spectral images of ocean surfaces using standard rgb images and typical ocean spectrum samples," *Optical Engineering*, vol. 60, no. 12, p. 123106, 2021. [Online]. Available: Available:https://doi.org/10.1117/1.OE.60.12.123106

[49] J. A. R. Miguel Antonio Barbero-Álvarez, José Manuel Menéndez, "An adaptive colour calibration for crowdsourced images in heritage preservation science," *IEEE Access*, vol. 8, pp. 185 093–185 111, 2020. [Online]. Available: Available:https://doi.org/10.1109/ACCESS.2020.3017390

[50] C. S. S. Prashant Saurabh Minz, "Comparison of computer vision system and colour spectrophotometer for colour measurement of mozzarella cheese," *Applied Food Research*, vol. 1, no. 2, p. 100020, 2021. [Online]. Available: Available:https://doi.org/10.1016/j.afres.2021.100020

[51] X. G. Tsai Cheng-Min, "A colour analysis tool for product appearance," *in E3S Web of Conferences*, vol. 179, pp. 1–5, 2020. [Online]. Available: Available:https://doi.org/10.1051/e3sconf/202017902064

[52] R. R. Tianfeng Chai, "Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature," *Geoscientific Model Development Draxler*, vol. 7, no. 3, pp. 1247–1250, 2014. [Online]. Available: Available:https://doi.org/10.5194/gmd-7-1247-2014

[53] P. K. Wan Sieng Yeo, Agus Saptoro, "Adaptive soft sensor development for non-gaussian and nonlinear processes," *Industrial Engineering Chemistry Research*, vol. 58, no. 45, pp. 20 680–20 691, 2019. [Online]. Available: Available:https://doi.org/10.1021/acs.iecr.9b03821

[54] W. S. Y. Teck Fu Thien, "A comparative study between pcr, plsr, and lw-pls on the predictive performance at different data splitting ratios," *Chemical Engineering Communications*, vol. 209, no. 11, pp. 1439–1456, 2021. [Online]. Available: Available:https://doi.org/10.1080/00986445.2021.1957853

[55] N. A. B. Wan Sieng Yeo, Mieow Kee Chan, "Prediction of glucose concentration hydrolysed from oil palm trunks using a plsr-based model," *in Intelligent Computing Optimization*, vol. 371, pp. 927–937, 2021. [Online]. Available: Available:https://doi.org/10.1007/978-3-030-93247-3_88

[56] A. S. Wan Sieng Yeo and P. Kumar, "Missing data treatment for locally weighted partial least square-based modelling: A comparative study," *Asia-Pacific Journal of Chemical Engineering*, vol. 15, no. 2, pp. 1–13, 2020. [Online]. Available: Available:https://doi.org/10.1002/apj.2422

[57] M. R. M. A. B. Y. Z. Y. C. L. L. G. K. S. V. N. Md Pervez, Wan Sieng Yeo, "Prediction of the diameter of biodegradable electrospun nanofiber membranes: An integrated framework of taguchi design and machine learning," *Journal of Polymers the Environment*, vol. 8, pp. 1–17, 2023. [Online]. Available: Available:https://doi.org/10.1007/s10924-023-02837-7

[58] K. M. Cort J Willmott, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005. [Online]. Available: Available:https://doi.org/10.3354/cr030079

[59] L. L. X. X. V. N. Y. C. Md Nahid Pervez, Wan Sieng Yeo, "Optimization and prediction of the cotton fabric dyeing process using taguchi design-integrated machine learning approach," *Scientific Reports*, vol. 13, no. 1, p. 12363, 2023. [Online]. Available: Available:https://doi.org/10.1038/s41598-023-39528-1

[60] P. K. M. K. Wan Sieng Yeo, Agus Saptoro, "Just-in-time based soft sensors for process industries: A status report and recommendations," *Journal of Process Control*, vol. 128, p. 103025, 2023. [Online]. Available: Available:https://doi.org/10.1016/j.jprocont.2023.103025

[61] R. D. Cramer, "Partial least squares (pls): its strengths and limitations," *Perspectives in Drug Discovery and Design*, vol. 1, no. 2, pp. 269–278, 1993. [Online]. Available: Available:https://doi.org/10.1007/BF02174528

[62] T. S. Hyojin Jung, "Comparison between the color properties of whiteness index and yellowness index on the cielab," *Textile Coloration and Finishing*, vol. 25, no. 4, pp. 241–246, 2013. [Online]. Available: Available:https://doi.org/10.5764/TCF.2013.25.4.241

[63] K. W. X. Y. Chenliang Liu, Yalin Wang, "Deep learning with nonlocal and local structure preserving stacked autoencoder for soft sensor in industrial processes," *Engineering Applications of Artificial Intelligence*, vol. 104, p. 104341, 2021. [Online]. Available: Available:https://doi.org/10.1016/j.engappai.2021.104341

[64] A. P. S. F. Aimi Aznan, Claudia Gonzalez Viejo, "Computer vision and machine learning analysis of commercial rice grains: A potential digital approach for consumer perception studies," *Sensors*, vol. 21, no. 19, p. 6354, 2021. [Online]. Available: Available:https://doi.org/10.3390/s21196354

[65] J. A. M. Lorri L Creech, "Ultraviolet radiation exposure in children: a review of measurement strategies," *Annals of Behavioral Medicine*, vol. 19, no. 4, pp. 399–407, 1997. [Online]. Available: Available:https://doi.org/10.1007/BF02895159

[66] W. S. Yeo, "Yellowness index prediction using locally weighted kernel partial least square regression model," *AIP Conference Proceedings*, vol. 2847, no. 1, p. 050001, 2023. [Online]. Available: Available:https://doi.org/10.1063/5.0165220

Wan Sieng Yeo is a senior lecturer in the Department of Chemical and Energy Engineering at Curtin University Malaysia. She was awarded a Doctorate and an MPhil. Engineering Degree in Chemical Engineering from Curtin University and a Bachelor's degree in Bioprocess-Chemical Engineering from Universiti Teknologi Malaysia. She also built up the skills and experiences from her engineering careers as a Technical Support Engineer, an Environmental Engineer, and a Trainee Chemical Engineer in both environmental and oil & gas industries before entering the academic field. She has also continued to gain her professional values and recognition via more than 10 awards, and 10 professional memberships that must complete the required Continuing Professional Development, CPD hours annually to develop lifelong learning. During her employment at Curtin University Malaysia, she has contributed to her roles and responsibilities through teaching, research, and service and leadership that aligns with the vision and mission of the university.

Agus Saptoro is a Professor at the Department of Chemical and Energy, Curtin Univesity Malaysia. He received a Bachelor of Engineering (cum laude) in chemical engineering from Gadjah Mada University Indonesia and obtained his Ph.D. in chemical engineering from Curtin University Australia. His research expertise and interest include multivariate statistics for data and image analyses, data-driven fault detection, diagnosis and modelling, inferential sensors, process modeling, simulation and control, and machine learning for process engineering. Agus is a Chartered Professional Engineer (CPEng) with Engineers Australia and an Executive member of the Malaysian Process Control Society (MyPCS).