Transferring Near Infrared Spectroscopic Calibration Model Across Different Harvested Seasons Using Joint Distribution Adaptation

Nur Aisyah Syafinaz Suarin and Kim Seng Chia(✉)

Department of Electronic Engineering, Faculty of Electrical
and Electronic Engineering, Universiti Tun Hussein Onn Malaysia,
86400 Parit Raja, Batu Pahat, Johor, Malaysia
kschia@uthm.edu.my

Abstract. Near infrared spectroscopic (NIRS) data from different harvested seasons may consist of different feature spaces even though the samples have the same label values. This is because the spectral response could be affected by the changes in environmental parameters, internal quality, and the reproducibility of NIRS instruments. Thus, this study aims to investigate the ability of Joint Distribution Adaptation (JDA) transfer learning algorithm in addressing the assumption of traditional machine learning i.e. both training and testing data must come from the same feature spaces and data distribution. First, NIRS data acquired from two different harvested seasons were used as the source domain and the target domain, respectively. Next, JDA was implemented to produce an adaptation matrix using the source domain and transfer datasets. This adaptation matrix would be used to transform the source and target domain datasets. After that, a calibration model was developed by means of Partial Least Squares (PLS) using the transformed training dataset; and validated using the transformed independent testing dataset. The proposed JDA-PLS was compared to the PLS without transfer learning as the baseline learning. Findings show that the proposed JDA-PLS with 10 LVs achieved the lowest RMSEP of 1.134% and the highest R_P^2 of 0.826.

Keywords: Transfer learning · Joint distribution adaptation · Near-infrared spectroscopy · Different seasons

1 Introduction

Near infrared spectroscopic (NIRS) is a promising non-destructive and fast technique to evaluate the quantitative and qualitative of organic materials. Electromagnetic energy of near infrared (NIR) spectrum is in a range of 750–2500 nm. The band range is known as one of the high-energy vibrational spectroscopy [1]. Interaction of emitted infrared radiation energy with samples is based on the chemical composition and physical properties of the samples [2]. Essentially, the samples examined by the NIRS consist of chemical bonds, i.e. C-H, N-H, C = O, that are able to absorb NIR energy [2]. In order to acquire high-quality NIR spectra from the examined samples, selecting the appropriate measurement setup is compulsory as the spectra are affected by the

N. A. Wahab and Z. Mohamed (Eds.): *Control, Instrumentation and Mechatronics: Theory and Practice*, LNEE 921, pp. 707–716, 2022.
https://doi.org/10.1007/978-981-19-3923-5_61

setup [3]. Usually, the selection of the setup is determined by the physical state of the samples, e.g. reflection is preferable for solid samples [3]. In earlier studies of NIR electromagnetic radiation, there was a long delay between the year it was discovered until it found the first analytical application [4]. This is due to the fact that the bands in NIR spectra are broad, correlated, and highly overlapping which required the presence of mathematical tools to extract the analytical information from these featureless spectra. Nowadays, with the progressive evolution in the NIRS instrument and mathematical resources, NIRS has been successfully explored and widely applied especially in agriculture and recently it has successfully contributed to the post harvested decision support system [5–7].

However, it is a tedious process, time-consuming, cost-intensive, and labour-intensive to establish NIRS predictive model [8]. A big dataset of targeted NIR spectra acquired from NIR data acquisition and chemical analysis needs to be collected and processed. Unfortunately, NIR spectra of the same sample collected from one NIRS device are different from another NIRS device, spectra collected from the same device of the same type with the same target value, but different populations also give out different spectra. The inconsistencies might be due to the measurement environmental conditions, samples internal qualities, or comes from the manufacturing process of NIRS instrument [9, 10]. Instead of performing calibration on a local dataset, calibration by using a global dataset had been proposed [11]. The result showed a better performance than using the local dataset, but the solution had high computational cost to establish the model, time-consuming to form the global dataset, and expensive in data collection. Thus, there is an urgent need for a robust and reliable model that can make a benefit from existing collected data to evaluate future data coming from different sources [12].

Calibration transfer is one of the renowned chemometrics techniques in transferring a calibration model between different spectrometers or generally is known as different domains. Calibration transfer is performed by transferring the model calibrated using a primary instrument to a secondary instruments using calibration transfer algorithm i.e. Direct Standardization (DS). Nevertheless, the calibration transfer suffers from the availability of standard samples with the same chemical constituents over time [13]. The reliance on standard samples for standardization remains a critical challenge for on-site applications as instruments are not always in the same location and recalibration needs sufficient labelled samples in each spectrometer. Hence, the calibration transfer approach is not convenient as a long-term problem-solving solution.

Recently, transfer learning (TL) has greatly improved the performance of many real-world applications in computer imaging and natural language processing [14–16]. The needs of TL occur when there was a limited labelled target domain dataset; while the availability of a related source domain dataset is sufficient to establish a learning model. Thus, the ability of TL to utilize knowledge present in labelled training data from a source domain to enhance a model performance in a target domain may be an alternative to address the limitation of calibration transfer. However, the studies of transfer learning for NIRS is limited in numbers; and thus more studies in constructing an efficient model using transfer learning are much needed [6, 17]. Thus, this study aims to evaluate and analyse the performance of transferred models from different domain feature spaces (across different harvest seasons) using transfer learning approach. Joint distribution adaptation (JDA) based Partial Least Squares (PLS) regression is proposed to evaluate the performance of dry matter content (DMC) of NIR mango predictive model across different harvested seasons.

2 Materials and Methods

2.1 Experimental Dataset

The effectiveness of the proposed method was evaluated through extensive experiments on mango dataset. The dataset was acquired and provided by Anderson et al. 2020 using a portable F750 Produce Quality Meter (Felix Instruments, Camas, USA) for the non-destructive NIR measurements; and an oven drying (UltraFD1000, Ezidri, Beverley, Australia) for dry matter content (DMC) measurement [9]. The mango dataset consisted of 11691 NIR spectra (684 – 990 nm) and DMC that measured from 4675 mango fruits from four harvested seasons in 2015, 2016, 2017, and 2018 [9, 18].

In this study, the mango dataset harvested in season 1 and season 4 were used. This is because this study aims to focus on transferring knowledge from the past season to recent season. Season 1 (i.e. 2015) was selected as the past season due to the distribution type of dataset in season 1 only coming from hard green dataset; while season 4 (i.e. 2018) consists of hard green and ripen type datasets. Thus, we can investigate the ability to transfer knowledge across harvested seasons with the different types of fruits. Table 1 summarises the information of the source domain and the target domain, and the data distribution for modelling the prediction model.

Table 1. Information of source domain, target domain and data distribution [9]

		Source domain	Target domain
Domain feature space, D = {X, P(X)}		Season 1	Season 4
Task, T = {Y, f(x)}		DMC	DMC
Year harvested		2015–2016	2018–2019
Sample size		3914	1448
Range of DMC value (min. – max.)		9.47–22.95	9.87–23.86
Type	Hard Green	3914	560
	Ripen	0	888
Number of samples for calibration		3914	0
Number of samples for transfer		0	14
Number of samples for validation		0	58
Number of samples for testing		0	1376

There were 3914 NIR spectra from season 1 and 1448 NIR spectra from season 4. Season 1 was fixed as the source domain; while season 4 was used as the target domain. All the dataset in season 1 was used as the training dataset; 5% of the earliest harvested mango dataset in season 4 was used as a transfer samples and validation dataset; while 95% of season 4 was used as the independent testing dataset.

2.2 Preprocessing of NIR Spectra and t-SNE Visualization

Spectra pre-processing is essential to eliminate or reduce phenomena such as background interference and instrument noises that existed in NIR spectra. Various preprocessing methods are available and one of the most common is standard normal variate (SNV) pre-treatment. The main purpose of data pre-processing is to reduce the complexity of spectra interpretation before the calibration process and to improve the accuracy of predicted models. In this study, three pre-processing were investigated to study the compatibility with the transfer learning approach, i.e. SNV, Savitzky-Golay (SG) and multiplicative scatter correction (MSC). SNV improves light scattering, MSC is able to correct linear effects and wavelength-dependent variations, and SG can eliminate spectrum baseline drift and reduce high-frequency noises. For the SG algorithm, the result can be affected by the selected window width. Useful information might be lost if the spectral distortion becomes severe due to the number of data points in the window. However, the smoothness of the denoising result is not ideal if the number of data points is too small. In this study, a second derivative based on a second-order polynomial across 17 points (SG2nd17) over the wavelength range 684−990 nm was adopted from previous work for comparison [9].

The relationship of DMC and 103 inputs per NIR spectra of mango harvested in season 1 and season 4 was visualized using the t-distributed Stochastic Neighbour Embedding (t-SNE) algorithm. t-SNE visualization algorithm was deployed to map the 103-dimensional NIR features into two-dimensional spaces using random walks on neighbourhood graphs. t-SNE is a more preferable tool to visualize high dimensional as compared with Principle Component Analysis (PCA) algorithm. This is because it could retain the non-linear structure of data to the maximum extent [19].

2.3 JDA Based PLS

Transfer learning is addressing the challenge of the assumption of traditional machine learning that training data and testing data must come from the same feature spaces and same data distribution. However, minimizing the related but different distribution of source and target domain dataset is difficult. Joint distribution adaptation (JDA) is the kind of feature-based transfer learning method, which jointly adapt both the marginal and conditional distributions in a principled dimensionality reduction procedure [20, 21]. Principal Component Analysis (PCA) is integrated with nonparametric Maximum Mean Discrepancy (MMD) in order to generate a new feature representation that closely matches the target domain. A detailed theoretic explanation about JDA can be found in a research article developed by Long et al. [20].

The objective of JDA is to find an orthogonal adaptation matrix such that the difference in both marginal and conditional distributions are minimized. The objective function is defined in Eq. (1).

$$\min_{A^T XHX^T A} \sum\nolimits_{c=0}^{C} tr\left(A^T XM_C X^T A\right) + \lambda ||A||_2^2 \tag{1}$$

where $X = [x_1, x_2, \ldots x_n] \in R^{mxn}$ is the input data matrix, $H = I - \frac{1}{n} l$ is the centre matrix, where $n = n_s + n_t$ and l is the nxn matrix of ones. n_s, n_t represent the number of samples in source domain and target domain, respectively. M is the original features number, λ is the regularization coefficient, C is the number of categories in regression problem. The MMD matrixes M_C computed as stated in Eq. (2):

$$(M_C)_{i,j} = \begin{cases} \frac{1}{n_s^c n_s^c} & x_i, x_j \in D_s^c \\ \frac{1}{n_t^c n_t^c} & x_i, x_j \in D_t^c \\ -\frac{1}{n_t^c n_t^c} & \begin{cases} x_i \in D_s^c, x_j \in D_t^c \\ x_j \in D_s^c, x_i \in D_t^c \end{cases} \\ 0 & otherwise \end{cases} \tag{2}$$

where, $D_s = \{(x_1, y_1), \ldots, (x_n, y_n\}$ is the labeled source domain, $D_t = \{(x_{ns+1}), \ldots, (x_{ns+nt})\}$ is the unlabeled target domain. Once the adaption matrix A is obtained, new feature representation $Z_{n_s} = A.x_{ns}$ with labelled DMC mango can be used to train a predictive model for target domain and make prediction with input as $Z_{n_t} = A.x_{n_t}$.

There are few hyperparameters in JDA need to be considered for a tuning in order to optimize the model performance. The hyperparameters are subspace bases, k and the regularization coefficient, λ. Grid search was implemented to structurally search the optimal values. The algorithm is searching the optimal solution in a search space by providing the minimum and maximum range for each hyperparameters. Table 2 shows the hyperparameter tuning search space for JDA based PLS.

Table 2. Hyperparameter tuning search space for JDA based PLS algorithm.

Hyperparameter	Initial implemented value	Search space [min.: max.]
Subspace bases, K	20	[20: 100]
Regularization coefficient, λ	0.1	[0.1: 1.0]

In the field of chemometrics in spectroscopy, PLS regression is one of the commonly applied calibration model. For the PLS regression algorithm, only one parameter needs to be tuned i.e. number of Latent Variables (LVs). Thus, PLS was used as the baseline regression (i.e. standard learning) in this research study. The number of LVs was chosen based on the minimum value of root mean squares error validation (RMSEV) of the established model. Figure 1 shows the comparison of basic workflow for standard learning and transfer learning for the general type of data. The difference between the two types of learning is the step where the transfer learning process takes place in the transfer learning – JDA based PLS before the calibration process. This is an important step in order to analyse the information and minimize the difference between source and target domain.

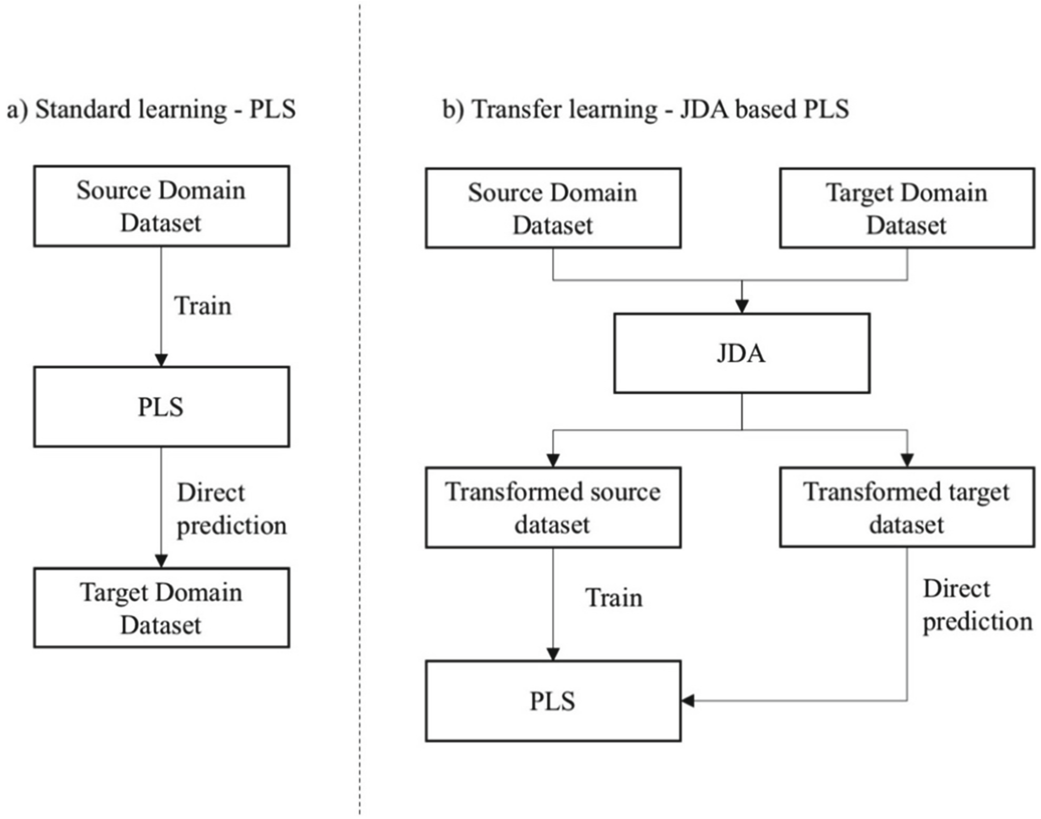


Fig. 1. Comparison of a) standard learning and b) transfer learning, the general process.

3 Model Evaluation

Statistical measurements root mean squared error (RMSE) and the coefficient of determination (R^2) are commonly used in NIRS model evaluation. A good NIRS model shall have a lower RMSE and higher R^2. Furthermore, to reflect the generalization ability of the model, the performance of the NIRS model was evaluated using independent training and testing datasets from different harvest seasons. The smaller the RMSE, the closer the simulated value to the measured value.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_p - y_i\right)^2}{n}} \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_p - y_i\right)^2}{\sum_{i=1}^{n}\left(y_p - \overline{y_i}\right)^2} \tag{4}$$

where y_i and y_p are the predicted and actual values of DMC mango, respectively. $\overline{y_i}$ is the mean of the DMC mango, and n is the sample size.

4 Results and Discussion

t-SNE was performed to visually examine the feature space of NIRS spectra of the source domain and target domain datasets. As shown in Fig. 2 , the clusters were formed by the DMC level in each dataset. The target domain dataset (season 4), shows the two obvious clusters that indicate two different types of mango, i.e. hard green and ripen. The distribution of dataset season 1 and season 4 were overlapped with each other for some of the data. In this way, it will be questioned whether the proposed approach would be able to capture the common features between the source domain and target domain. Besides, the graph plotting of t-SNE further demonstrated the suitability of DMC prediction by using the NIR technique.

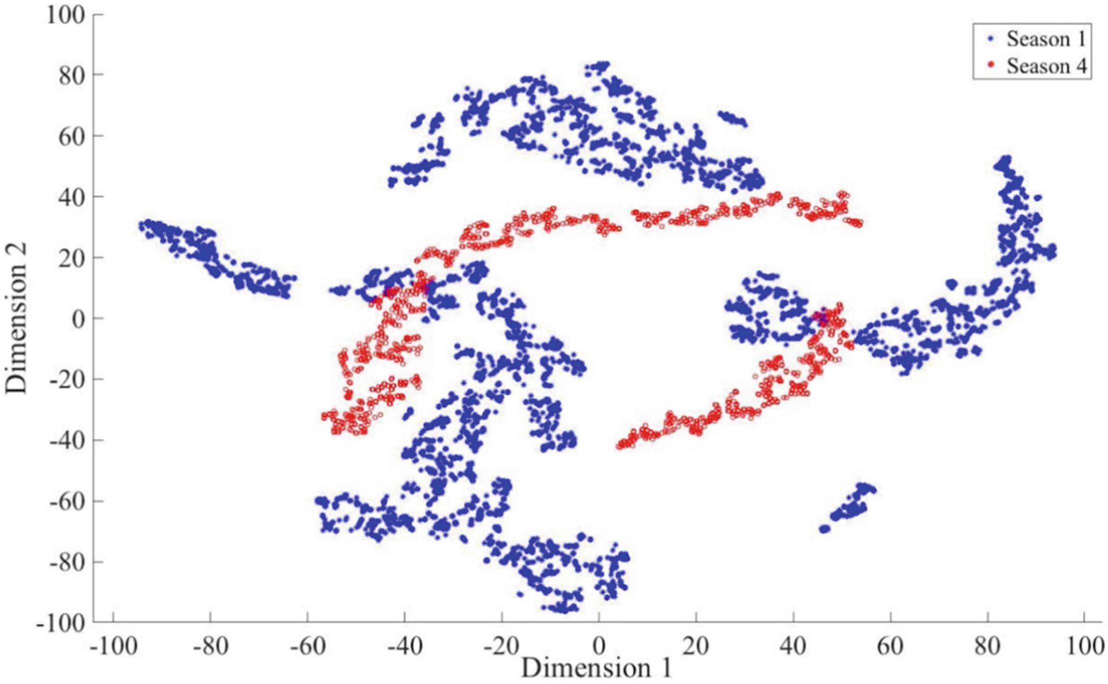


Fig. 2. t-SNE visualization of source domain – Season1 (blue) and target domain – Season 4 (red) feature spaces.

Table 3 tabulates the performance of DMC predictive models for NIRS mango. The proposed JDA based PLS obtained the lowest RMSEP and the highest R_P^2. This proves the ability of JDA to minimize the difference of different feature spaces between season 1 and season 4. Furthermore, by only using the transformed season 1 dataset, JDA based PLS successfully enhanced the performance of standard baseline learning of PLS. In other words, JDA has addressed the limitation of PLS to sustain its robustness across different harvest seasons that have different feature distributions.

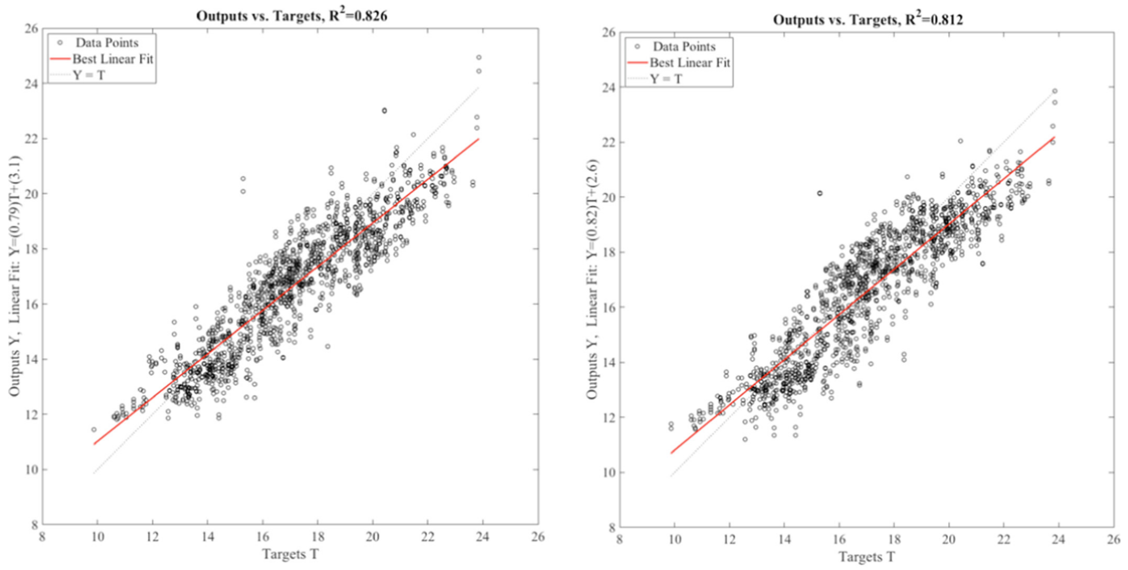
For each learning approach, the PLS model that developed using different pre-processing methods were included and compared to analyze the most appropriate pre-processing method for the mango dataset with the different learning methods. When the NIR spectra were pre-processed by MSC, SNV, and SG2nd pre-processing algorithms, JDA based PLS were degraded. This could be due to there were different noises between season 1 and season 4. Season 4 consisted of ripen type of mango samples. According to Seifert et al., scattering properties of fruit would change during maturation process [22]. Thus, hard green and ripen fruit could have different scattering effects. As a result, applying a same pre-processing method to both datasets would degrade the model performance, compared with untreated NIR spectra. In other words, applying an appropriate pre-treatment to the NIR spectra is important to preserve the information in the spectra.

Meanwhile, for the performance of baseline learning, the PLS that used 14 LV's with SG2nd17 treatment shows the lowest RMSE of 1.555%, and the highest R^2 of 0.812. This is aligned with the result that published by Ander-son et.al as SG2nd17 was adapted from the previous study [9]. However, the model required the highest number of LV than others as the input has become more complicated to establish good predictive model.

Table 3. Prediction model performances for NIRS mango DMC using transfer learning approach (JDA based PLS) and standard learning approach (PLS).

	Treatment	LV	RMSEC	R_c^2	RMSEV	R_V^2	RMSEP	R_P^2
Transfer learning: JDA based PLS	**RAW**	**10**	**0.982**	**0.858**	**0.723**	**0.885**	**1.139**	**0.826**
	SNV	9	1.124	0.794	0.963	0.926	1.628	0.783
	MSC	9	1.183	0.814	0.963	0.926	1.628	0.783
	SG2nd17	8	1.282	0.723	0.469	0.951	1.715	0.694
Baseline method: PLS	RAW	10	1.006	0.851	1.006	0.851	1.612	0.794
	SNV	9	1.153	0.804	1.153	0.804	1.654	0.722
	MSC	8	1.313	0.746	1.313	0.746	1.705	0.642
	SG2nd17	14	0.719	0.920	0.719	0.920	1.555	0.812

Figure 3(a) and (b) show the regression of JDA based PLS that used 10 LVs without pretreatment; and that of baseline PLS that used 14 LVs with SG2nd17 pretreatment, respectively. For Fig. 3(a), 10 LVs were chosen based on the lowest validation result using the sample transfer dataset. The model reached RMSEP of 1.139% when it was tested using the independent testing dataset (i.e. NIR mango harvested in season 4). The RMSEP was drastically increased to 1.555% when the same dataset has established a model using the baseline method, PLS without a transfer learning approach. This shows that the generalization of the PLS was poor when it was applied directly to another harvest season. When JDA transfer learning algorithm was integrated, JDA based PLS achieved a lower RMSEP compared to that without JDA on the different harvested seasons. This could be due to NIR spectral responses of an organic sample are sensitive to the changes of the environment. Consequently, this shows that the use of JDA was able to match or minimize the differences among different domain distributions and to improve the robust of the predictive model across different harvest seasons.



(a) (b)

Fig. 3. Performance of (a) JDA based PLS model that used 10 LVs without pretreatment; (b) PLS that used 14 LVs with SG2nd17 pretreatment

5 Conclusion

NIRS model may need to be recalibrated for NIRS data that have different distribution of a similar type of samples. In this study, JDA based PLS (JDA-PLS) was introduced and evaluated using two NIRS mango datasets that acquired from different harvested seasons. The experimental results showed that the proposed JDA-PLS has a positive result in transferring knowledge across different domain distributions (i.e. across different harvest seasons). Results show that the proposed JDA-PLS achieved the lowest RMSEP of 1.139% and the highest R_P^2 of 0.826. This result demonstrated that the proposed JDA-PLS transfer learning approach is promising to overcome the concern of NIRS models across different population samples that have different distributions. Thus, this study suggests that the proposed solution can be further explored and utilized for different cases and experimental design of transferring knowledge across NIR samples as transfer learning for NIR still limited.

Acknowledgement. This research was supported by Ministry of Higher Education (MOHE) through Fundamental Research Grant Scheme (FRGS) FRGS/1/2020/ICT02/UTHM/03/6. The authors would also like to acknowledge Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia for providing facilities for this study.

References

1. Pasquini, C.: Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. J. Braz. Chem. Soc. **14**, 198–219 (2003)
2. Junior, S.B., et al.: Multi-target prediction of wheat flour quality parameters with near infrared spectroscopy. Inf. Process. Agric. **7**, 342–354 (2020)

3. Hong, F.W., Chia, K.S.: A review on recent near infrared spectroscopic measurement setups and their challenges. Meas. J. Int. Meas. Confed. **171**, 108732 (2021)
4. Pasquini, C.: Near infrared spectroscopy: a mature analytical technique with new perspectives – a review. Anal. Chim. Acta. **1026**, 8–36 (2018)
5. Walsh, K.B., McGlone, V.A., Han, D.H.: The uses of near infra-red spectroscopy in postharvest decision support: a review. Postharvest Biol. Technol. **163**, 111139 (2020)
6. Yu, Y., Huang, J., Zhu, J., Liang, S.: An accurate noninvasive blood glucose measurement system using portable near-infrared spectrometer and transfer learning framework. IEEE Sens. J. **21**, 3506–3519 (2021)
7. Goldshleger, N., Grinberg, A., Harpaz, S., Shulzinger, A., Abramovich, A.: Real-time advanced spectroscopic monitoring of Ammonia concentration in water. Aquac. Eng. **83**, 103–108 (2018)
8. Mishra, P., Woltering, E., Brouwer, B., Hogeveen-van Echtelt, E.: Improving moisture and soluble solids content prediction in pear fruit using near-infrared spectroscopy with variable selection and model updating approach. Postharvest Biol. Technol. 171, 111348 (2021)
9. Anderson, N.T., Walsh, K.B., Subedi, P.P., Hayes, C.H.: Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content. Postharvest Biol. Technol. **168**, 111202 (2020)
10. Mishra, P., Passos, D.: Realizing transfer learning for updating deep learning models of spectral data to be used in new scenarios. Chemom. Intell. Lab. Syst. **212**, 104283 (2021)
11. Yap, X.Y., Chia, K.S.: A comparison between local and global models among different near infrared spectroscopy instruments for corn oils prediction. In: Proceeding - 2021 IEEE 17th International Colloquium on Signal Processing and its Applications CSPA 2021, pp. 111–115 (2021). https://doi.org/10.1109/CSPA52141.2021.9377295
12. Mishra, P., et al.: Are standard sample measurements still needed to transfer multivariate calibration models between near-infrared spectrometers? The answer is not always. TrAC - Trends Anal. Chem. **143**, 116331 (2021)
13. Chen, Y.Y., Wang, Z.B.: Cross components calibration transfer of NIR spectroscopy model through PCA and weighted ELM-based TrAdaBoost algorithm. Chemom. Intell. Lab. Syst. **192**, 103824 (2019)
14. Baydilli, Y.Y., Atila, U., Elen, A.: Learn from one data set to classify all – A multi-target domain adaptation approach for white blood cell classification. Comput. Methods Programs Biomed. 196, (2020)
15. Zhao, K., Jiang, H., Wang, K., Pei, Z.: Joint distribution adaptation network with adversarial learning for rolling bearing fault diagnosis. Knowledge-Based Syst. **222**, 106974 (2021)
16. Farahani, A., Pourshojae, B., Rasheed, K., Arabnia, H.R.: A Concise Review of Transfer Learning. Proc. - 2020 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2020. 344–351 (2020)
17. Qiu, Z., Zhao, S., Feng, X., He, Y.: Transfer learning method for plastic pollution evaluation in soil using NIR sensor. Sci. Total Environ. **740**, 140118 (2020)
18. Mishra, P., Passos, D.: Deep chemometrics: Validation and transfer of a global deep near-infrared fruit model to use it on a new portable instrument. J. Chemom. 1–12 (2021)
19. Pahar, M., Klopper, M., Warren, R., Niesler, T.: COVID-19 Detection in Cough, Breath and Speech using Deep Transfer Learning and Bottleneck Features. Comput. Biol. Med. 105153 (2021)
20. Long, M., Wang, J., Ding, G., Sun, J., Yu, P.S.: Transfer feature learning with joint distribution adaptation. Proc. IEEE Int. Conf. Comput. Vis. 2200–2207 (2013)
21. Liu, W., Liu, W.D., Gu, J.: Predictive model for water absorption in sublayers using a Joint Distribution Adaption based XGBoost transfer learning method. J. Pet. Sci. Eng. **188**, 106937 (2020)
22. Seifert, B., Zude, M., Spinelli, L., Torricelli, A.: Optical properties of developing pip and stone fruit reveal underlying structural changes. Physiol. Plant. **153**, 327–336 (2015)