https://doi.org/10.31436/iiumej.v25i1.2796

SPECTROSCOPY DATA CALIBRATION USING STACKED ENSEMBLE MACHINE LEARNING

MAHMUD IWAN SOLIHIN[1*], CHAN JIN YUAN[1], WAN SIU HONG[1], LIEW PHING PUI[2], CHUN KIT ANG[1], WAFA HOSSAIN[1] AND AFFIANI MACHMUDAH[3]

[1]*Faculty of Engineering, Technology and Built Environment, UCSI University, Malaysia*
[2]*Faculty of Applied Sciences, UCSI University, Malaysia*
[3]*Research Centre for Hydrodynamics Technology, National Research and Innovation Agency (BRIN), Surabaya, Indonesia*

[*]*Corresponding author: mahmudis@ucsiuniversity.edu.my*

(Received: 20 March 2022; Accepted: 4 June 2023; Published on-line: 1 January 2024)

ABSTRACT: Near infrared spectroscopy (NIRS) is a widely used analytical technique for non-destructive analysis of various materials including food fraud detection. However, the accurate calibration of NIRS data can be challenging due to the complexity of the underlying relationships between the spectral data and the target variables of interest. Ensemble learning, which combines multiple models to make predictions, has been shown to improve the accuracy and robustness of predictive models in various domains. This paper proposes stacking ensemble machine learning (SEML) for calibration of NIRS data with two levels of learning involved. Eight (8) spectroscopy datasets from public repository and previously published works by the authors are used as the case study. The model well generalized the data in the respective regression tasks with R^2 of at least ≈0.8 in the test samples and in the respective classification tasks with classification accuracy (CA) of at least ≈0.8 also. In addition, the proposed SEML can improve, or at least reach par with, the accuracy of individual base learners in both train and test samples for all cases of regression and classification datasets. It shows superior performance in test samples for both regression and classification datasets with respectively R^2 ranging from 0.86 to nearly 1 and CA ranging from 0.89 to 1.

ABSTRAK: Spektroskopi inframerah dekat (NIRS) adalah teknik analitikal yang banyak digunakan bagi analisa pelbagai bahan tanpa merosakkan bahan termasuk ketika mengesan penipuan makanan. Walau bagaimanapun, kalibrasi yang tepat bagi data NIRS adalah sangat mencabar kerana hubungan antara data spektral dan pemboleh ubah sasaran yang ingin dikaji bersifat kompleks. Gabungan pembelajaran (*Ensemble learning*), iaitu gabungan pelbagai model bagi membuat prediksi, telah terbukti dapat meningkatkan ketepatan dan kecekapan model prediksi dalam pelbagai bentuk. Kajian ini mencadangkan Turutan Gabungan Pembelajaran Mesin (*Stacking Ensemble Machine Learning*) (SEML), bagi teknik penentu ukuran data NIRS melibatkan dua tahap pembelajaran. Lapan (8) set data spektroskopi dari repositori awam dan kajian terdahulu oleh pengarang telah digunakan sebagai kes kajian. Model ini menggeneralisasi data dalam tugas regresi R^2 masing-masing sebanyak H0.8 bagi sampel ujian dan pengelasan tugas masing-masing dengan ketepatan klasifikasi (CA) sekurang-kurangnya H0.8. Tambahan, SEML yang dicadangkan ini dapat membantu, atau sekurang-kurangnya setanding dengan ketepatan individu dalam pembelajaran berkumpulan dalam kedua-dua sampel latihan dan ujian bagi semua kes set data regresi dan klasifikasi. Ia menunjukkan prestasi terbaik dalam sampel ujian bagi kedua-dua kumpulan set data regresi dan klasifikasi dengan masing-masing R^2 antara 0.86 hingga hampir 1 dan antara julat 0.89 hingga 1 bagi CA.

KEYWORDS: *chemometrics calibration; stacking ensemble machine learning; near infrared spectroscopy; food fraud detection; food safety and security*

1. INTRODUCTION

Near infrared spectroscopy (NIRS) is a sort of high-energy vibrational spectroscopy that operates in the wavelength range of 750 to 2500 nm (13333 to 4000 cm^{-1}). By probing a sample with electromagnetic radiation in that wavelength range, NIRS obtains spectral information that can aid in the development of appropriate qualitative and/or quantitative analytical procedures. NIRS has gained wide acceptance of industrial applications and research as a secondary non-destructive material fingerprinting. The applications can include medical-pharmaceutical fields [1-3], food/agricultural analysis [4-7], forensic [8], etc. NIRS is described as the hallmark for one of the most rapidly advancing analytical techniques over the last few decades [9].

The calibration of NIR spectra into meaningful quantitative or qualitative information is normally performed using advanced statistical learning, or chemometrics, analysis. Chemometrics methods are used for analyzing molecular spectroscopy data such as near infrared (NIR), Fourier transform infrared (FTIR), ultraviolet–visible (UV-vis), induced breakdown spectroscopy (LIBS), Raman spectroscopy and nuclear magnetic resonance (NMR) spectroscopy [10]. This problem is essentially multivariable data analysis or calibration to reveal meaningful chemical information from the samples being scanned by the respective spectrometers.

Due to advancement of computational methods, recently machine learning (ML) and deep learning are also used in NIRS chemometrics analysis. Generally, their applications enhance the calibration accuracy performed using common conventional statistical learning such as partial least square (PLS) regression and linear discriminant analysis (LDA). For example, the study in [11] has used deep and ensemble learning for milk adulteration detection using Fourier transformed infrared spectroscopy (FTIR). The proposed neural network architecture can outperform the composition recognition made by commonly used methods. In another study [4], support vector machine (SVM) was used to perform regression on NIR spectra data of mango fruits brix level (sugar content). The SVM performance outperforms PLS algorithm. SVM was also used for food powder classification based on handheld NIR spectrometer data and the results were excellent [12]. Another study [13] also using SVM as a classifier to estimate the sample quality of Andrographis paniculate obtained from different sources. The NIR reflectance spectroscopy instrument was used here.

Furthermore, Chen et al. [14] proposed the Least Squares Support Vector Machine (LSSVM) algorithm to establish NIR calibration models for the quantitative determination of chemical oxygen demand, which is a critical indicator of water pollution level. Michel et al. [15] used k-nearest neighbors (KNN) in addition to SVM, combined with principal component analysis (PCA) for identifying type of both consumer plastics and marine plastic debris based on different spectroscopy data namely Fourier transform infrared spectroscopy (ATR−FTIR), NIR reflectance spectroscopy, laser-induced breakdown spectroscopy (LIBS), and X-ray fluorescence (XRF) spectroscopy. Success rates indicate that ATR−FTIR, NIR reflectance spectroscopy, and LIBS coupled with ML classifiers can be used to identify both consumer and environmental plastic samples. Perez et al. [16] showed good results for the classification of chicken meat parts, where the portable NIR spectrophotometer together with chemometrics and ML algorithms allowed to discriminate the different parts of chicken by LDA, random forest (RF), and SVM. Wang et al. [17] investigated three conventional ML methods, namely

https://doi.org/10.31436/iiumej.v25i1.2796

ordinary least square estimation (OLSE), RF, and extreme learning machine (ELM), while for the deep learning methods, three different structures of convolutional neural network (CNN) incorporated Inception module were constructed and investigated. The study conducted chemometrics calibration for total soil nitrogen using Visible-near-infrared spectrum (Vis-NIR) spectroscopy. The results indicate that the baseline-corrected and smoothed ELM model reached practical precision (coefficient of determination, R^2=0.89) and the best result of CNN was obtained with R^2=0.93.

ML is basically a data-driven modelling that can be applied for multivariable calibration like in NIR spectroscopy, i.e. chemometrics calibration. As it has been discussed, there are various ML algorithms applied for multivariable calibration methods. Ensemble ML emerges as an effort for improving prediction accuracy of individual ML models by combining multiple ML models into a final prediction output. RF is one of the best ensemble ML that performs well in many applications [18]. RF is basically a multiple model of decision tree. Another approach of ensemble ML is called stacking ensemble. In the stacking ensemble ML, some base ML algorithms will be used as first level (base) learners and logistic regression (LR) will be used as the second level learner to aggregate the outputs of first level learners and come up with the final prediction output [19].

However, the applications ensemble ML for NIR spectroscopy calibration still needs to be further explored. Particularly, to our knowledge, stacking ensemble ML applied for multiple NIRS datasets is not found in the literature. Most of research works applied the method for an indicidual NIRS data such in [20]. Our main contribution in this research is the implementation of stacking ensemble machine learning (SEML) for chemometrics calibration where multiple NIR spectroscopy datasets will be used as the case study involving classification (qualitative) and regression (quantitative). This is proposed to improve the accuracy of the calibration model using conventional calibration methods. The performance of the proposed stacking ensemble ML will be compared to the base ML algorithm and conventional statistical learning in chemometrics.

2. SPECTROSCOPY DATASETS

There are two types of NIR spectroscopy multivariable calibrations from the perspective of supervised ML, namely classification (class label/qualitative prediction) and regression (quantitative prediction). Here, four datasets are used from each type thus making a total of 8 NIR spectroscopy datasets. For the regressions problem, the four datasets are adulteration of honey (AH) [18], active substance in a pharmaceutical tablet (AST) [21], dry matter content within mango fruit (DMM) [22] and moisture content of grain protein (MGP) [23]. The four datasets used for the classification task are adulteration of rice dataset (RA) [6], coffee type (CF) dataset [24], strawberry fruit (SB) [25] and starch powder classification (SP) [12]. The summary of the datasets and their attributes are shown in Table 1 (Regression and Classification).

2.1 Regression Datasets

The following are the details of the four regression datasets. The first regression dataset is the AH dataset, which deals with the regression of the level of adulteration of Kelulut honey (Malaysia) with syrup. The adulteration levels are given from 0%, 10%, etc., up to 100%. The level of 0% adulteration means pure honey sample and vice versa. NIR spectra data were collected using a micro NIR handheld instrument with a wavelength range of 900-1700 nm. The data are described in more detail in the study in [18]. Note that the original wavelength points are from 900 nm to 1700 nm, but the data is cut at the longer wavelengths due to the

https://doi.org/10.31436/iiumej.v25i1.2796

presence of noise, resulting in data at wavelengths of 900 nm to 1650 nm. In this study [18], the calibration was performed in classification mode where k-nearest neighbour (KNN) and random forest were used, achieving accuracy of 90%. In this paper, we will extend this study and convert the problem into regression mode which has not been done before.

The second regression dataset deals with the chemometrics quantitation of the active substance in a pharmaceutical tablet (% per tablet), i.e. AST dataset. Four different dosage values of this pharmaceutical drug (5, 10, 15, and 20 mg per tablet) were used. In total, 31 batches were used, and from each batch 10 tablets were individually weighed and analyzed: making up 310 NIR spectra data. The data are described in more detail in the study in [21]. NIR transmittance and Raman spectroscopy chemometric calibrations of the active substance content of a pharmaceutical tablet were developed using partial least-squares regression (PLS) and no machine learning calibration were involved. The results gave relative prediction errors (RMSECV/ynom) of 2.6–3.7%. The latest study that employs this dataset was in 2021 [26] where KNN, SVM, RF and deep learning were used. However, the AST data was used in classification mode and not regression. We will use the AST dataset for the regression problem.

The third regression dataset deals with the dry matter content (%) regression of mango fruit. Let us call it the DMM dataset. Originally, the data for intact mango fruit was collected with short wave near infrared spectra using an interactance geometry, with total data set collected across three seasons (n = 10243) and that of a fourth season (n = 1,448) consisting of 306 wavelength points [22]. The dataset was reduced after pre-processing to have a number of samples (n=11362) with 103 wavelength points (features) as published in [27]. In the paper [22], PLS regression and ANN were used as regression models and they achieve similar performance with Root Mean Square Error of Prediction (RMSEP) of around 1%.

The fourth regression dataset deals with regression of the moisture (wt%) of grain protein from NIR instrument with 231 samples created by Tormod Naes and Tomas Isaakson, as described on the website [23]. We call it the MGP dataset. The NIR spectrum have 117 wavelength points, ranging from 1104 to 2495 nm. The latest publication that used this dataset was found in [28]. The reported calibration result was obtained with $R^2 = 0.93$. This still can be improved and there was no involvement of ML algorithms during calibration.

Table 1: Summary of the eight (8) spectroscopy datasets used in this study

Calibration types	Dataset	Target variable	Wavelength range	Number of samples	Instrument
Regression (quantitative)	AH	Adulteration level %) – kelulut honey	900-1700 nm	1846	Micro handheld NIR
	AST	Active substance (% per tablet) – tablet	7400-10500 cm^{-1} (~950 nm to 1350 nm)	309	NIR FT-Raman
	DMM	Dry matter content (% weight) – mango fruit	684-990 nm	11362	F750 handheld NIR
	MGP	Moisture (% weight) – grain protein	1100-2500 nm	231	NIR reflectance
Classification (qualitative)	RA	Authentic and adulterated rice	900-1700 nm	123	Micro handheld NIR
	CF	Two categories of coffee	810-1910 nm	56	FTIR spectroscopy
	SB	Strawberry and non-strawberry	900-1800 nm	983	FTIR spectroscopy
	SP	Five categories of flour	900-1700 nm	75	Micro handheld NIR

https://doi.org/10.31436/iiumej.v25i1.2796

2.2 Classification Datasets

Next, the four classification datasets are explained as follows. The first classification dataset is rice adulteration (RA dataset) where a total of 123 NIR spectra data were collected from 31 unadulterated rice samples and ten adulterated rice samples in 3 different illumination conditions. Rice samples were bought from Tesco Hypermarket and NSK Trade City, Kuala Lumpur, Malaysia. The rice samples were all milled rice in which the rice husk was removed. The total of 31 rice samples included 14 brands of local white rice, 10 brands of Thailand fragrance rice, 3 brands of Thailand white rice, and four other types of rice. Rice replica was bought from Titoonic Enterprise (Malaysia) and chosen as the adulterant [6]. For the chemometric calibration, Principal Component Analysis (PCA) and Logistic Regression (LR) were used simulatanoeusly to perform qualitative analysis whether the rice sample was adulterated on unadulterated, achieving 97.2% accuracy. This accuracy was only acieved with inclusion of PCA as feature reduction. We will extend the study using different machine learning algorithms and no PCA will be used to simplify the process.

The second classification dataset is the CF dataset which is to discriminate the samples according to two coffee species, robusta and arabica. A total of 56 spectra samples were acquired using DRIFT (diffuse reflection infrared Fourier transform) and ATR (attenuated total reflection) sampling techniques in FTIR spectroscopy. The dataset is publicly accessible from the website [24]. The most recent publication using this dataset was in [29] where deep learning convolutional neural networks (CNN) was used in the calibration resulting in 100% accuracy. From the computation persfective, deep learning is more complex than ML.

The third classification dataset is the SB dataset which deals with strawberry adulteration discrimination [24]. A total of 983 FTIR spectra data were collected representing two sample types, i.e., strawberry and non-strawberry puree. The adulterated strawberry puree samples were obtained by mixing with certain adulterants as discussed in the study. Recent publication using this dataset was found in [30]. The results of classification accuracy using FM (frequency modulation) synthesis as the sound synthesiser and PCA as the dimensionality reduction method yields a mean classification accuracy of 88.57%. The result can be improved, and a more advanced ML algorithm can be used, as will be done in this paper.

The fourth classification dataset is the SP dataset, which is used to differentiate between various types of starch powder, including organic wheat flour, non-organic wheat flour, rice flour, tapioca starch, and corn starch [12]. The NIR spectra data was collected using micro NIR spectroscopic instrument with wavelength range from 900-1700 nm. A total of 75 NIR spectra samples were collected for five different food powder types, i.e., 15 samples for each type. Here, SVM was used as ML algorithm for chemometric calibration producing 100% test accuracy. However, as SVM involve extensive kernel tuning, we will apply different ML algoritms to study the feasibility.

3. MACHINE LEARNING FOR CHEMOMETRICS CALIBRATION

The qualitative or quantitative information from infrared spectra data is only obtained after the calibration process using chemometrics and this process naturally involves multivariate statistical analysis. Machine learning (ML) as a subset of AI (artificial intelligence), in addition to conventional multivariate statistical tools, seems to get more popularity for chemometrics calibration in spectroscopy nowadays due to its well-known capability to perform complex classification and regression tasks based on the data provided, i.e. data-driven method [33].

https://doi.org/10.31436/iiumej.v25i1.2796

Regardless of their suitability of application for regression or classification, some famous ML algorithms are artificial neural networks (ANN), support vector machine (SVM), k-Nearest Neighbor (KNN), Naïve Bayes (NB), Gradient Boosting (GB), Random Forest (RF), etc. In this study, we focus on applying the ANN, SVM, GB and RF algorithms in the ensemble ML scheme. Among the considerations for choosing these ML algorithms are ease of interpretation and visualization, requiring less effort on data pre-processing, i.e., not requiring scaling and normalization, and not largely influenced by outliers or missing values.

Further improvement of standard ML model is called ensemble ML. The idea of ensemble ML is basically creating multiple ML models and aggregating the final prediction using a certain method to improve the accuracy of each individual base model performance. There are fundamentally three methods of ensemble namely bagging (bootstrap aggregating), boosting, and stacking.

Bagging typically involves using a single machine learning algorithm, almost always an unpruned decision tree (DT), and training each model on a different sample of the same training dataset. The predictions made by the ensemble members are then combined using simple statistics, such as voting or averaging [32]. A popular example of bagging ensemble method is random forest (RF) which is basically the bagging ensemble of DT.

Boosting on the other hand, refers to a family of algorithms that can convert weak learners to strong learners. Intuitively, a weak learner is just slightly better than random guess, while a strong learner is very close to perfect performance [19]. It involves the use of very simple decision trees that only make a single or a few decisions, referred to as 'weak learners'. The predictions of the weak learners are combined using simple voting or averaging, although the contributions are weighed proportional to their performance or capability. The objective of Gradient Boosting (GB) is to develop a so-called 'strong learner' from many 'weak learners'. Adaptive Boosting (Adaboost) and eXtreme Gradient Boosting (XGB) are among the popular methods in boosting ensemble [34]. The general boosting procedure is described in Fig. 1 [19].

Input: Data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
Base learning algorithm $\mathcal{L}$;
Number of learning rounds T.

Process:

1. $\mathcal{D}_1(x) = \mathcal{D}(x)$ % Initialize distribution
2. **for** $t = 1, \ldots, T$:
3. $h_t = \mathcal{L}(\mathcal{D}_t)$; % Train a weak learning learner h_t from distribution $\mathcal{D}_t$
4. $\epsilon_t = P_{x \sim D_t}\big(h_t(x) \neq f(x)\big)$; % Evaluate the error of h_t
5. $\mathcal{D}_{t+1}(x) = Adjust\_Distribution\ (\mathcal{D}_t, \epsilon_t)$
6. **end**

Output: $H(x) = Combine\_Outputs(\{h_1(x), \ldots, h_t(x)\})$

Fig. 1: General procedure of boosting algorithm.

4. STACKING ENSEMBLE PROCEDURE

Stacking ensemble ML (SEML) is proposed for spectroscopy data calibration in this paper. In stacking ensemble, two levels of learning are used where 1st level (base) learners can be different ML algorithms and a 2nd level learner is used to combine the predictions (normally a logistic regression for classification). In this study, the proposed stacking ensemble ML algorithm uses four ML algorithms as base learners (1st level learners) namely Gradient Boosting (GB), Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural

https://doi.org/10.31436/iiumej.v25i1.2796

Network (ANN). The 2nd level learner uses Logistic Regression (LogReg) for the case of classification and uses Linear Regression (LinReg) for the case of regression to aggregate the final output.

Thus, suppose that $\hat{y}_{reg}$ and $\hat{y}_{class}$ are the final prediction output for regression and classification respectively, they can be expressed as:

$$\hat{y}_{reg} = LinReg(\hat{y}_{GB}, \hat{y}_{RF}, \hat{y}_{SVM}, \hat{y}_{ANN}) \quad (1)$$

$$\hat{y}_{class} = LogReg(\hat{y}_{GB}, \hat{y}_{RF}, \hat{y}_{SVM}, \hat{y}_{ANN}) \quad (2)$$

Accordingly, the diagram of the proposed stacking ensemble, ML, is shown in Fig. 2. Here, the algorithm takes the process as illustrated in Fig. 3. The stacking ensemble learns a high-level classifier/regressor on top of the base learner. It can be regarded as a meta learning approach in which the base learners are called first-level learners and a second-level models have learnt to combine the first-level learners.

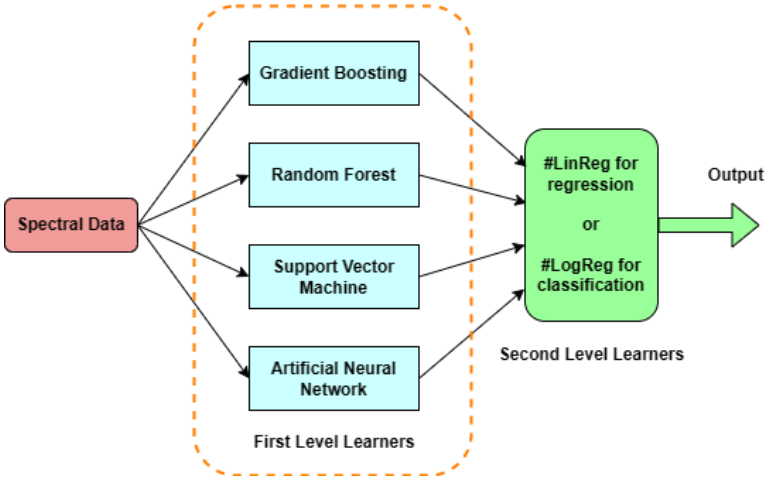

Fig. 2: Diagram of the stacking ensemble ML for regression and classification.

Input: Data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
Process:
1. Step 1: Learn first level learners
2. **for** $t = 1, \ldots, T$
3. Learn a base classifier/regressor h_τ based on D
4. **end**
5. Step 2: Construct new data sets from D
6. **for** $i = 1, \ldots, m$
7. Construct a new dataset that contains $\{x_i', y_i\}$, where $x_i' = \{h_1(x_i), h_2(x_i), \ldots, h_\tau(x_i)\}$
8. **end**
8. Step 3: Learn a second level classifier/regressor
9. Learn a new classifier/regressor h' based on the newly constructed dataset
Output: An ensemble model $H(x) = h'(h_1(x), h_2(x), \ldots, h_\tau(x))$

Fig. 3: General procedure of stacking ensemble.

Based on Fig. 3, the general procedure of stacking ensemble has the following three major steps:

https://doi.org/10.31436/iiumej.v25i1.2796

- Step 1: Learn first-level learners (GB, RF, SVM, ANN) based on the original training data set.
- Step 2: Construct a new data set based on the output of base learners. Here, the output predicted labels/values of the first-level learners are regarded as new features, and the original labels/values are kept as the labels/values in the new data set.
- Step 3: Learn a second-level learner based on the newly constructed data set. LinReg and or LogReg are applied as second-level learners for regression and classification tasks respectively.

5. CALIBRATION MODEL DEVELOPMENT AND EVALUATION

As the general procedure in ML model development, the ML is trained using a training dataset. Once the training is completed, the model is tested using the test dataset to evaluate the prediction accuracy. Here, 70% - 30% ratio is used to assign the size of the training and testing datasets, respectively. In the random assignment into the classification datasets, stratified sampling is applied to make sure a balanced class is achieved in the training and testing datasets. Spectra data preprocessing is also applied according to the common sense of spectra data processing experience. Simple preprocessing, such as edge cutting, Gaussian smoothing, and Savitsky-Golay (SG) derivative filter, is executed in this study to show that the proposed ML model is robust to preprocessing methods.

Table 2 shows the spectra pre-processing used and the ML hyperparameters setup of the individual base learners and the stacking ensemble ML. The implementation of the ML training and evaluation is performed in the Python programming environment. For the spectra preprocessing, edge cutting is necessary as the micro handheld NIR spectrometer (used in AH, BG, BM, SP, RA datasets) produced noise on both edges of the 900-1700 nm wavelength window. Here, edge cutting means to cut and keep the wavelengths at 950-1650 nm for these four datasets. For all datasets, gaussian smoothing is applied to reduce signal noise. For all classification datasets, SG derivative order 1 is applied to remove effects of shifting baselines and sloping or curving due to scattering [35]. For the regression datasets, SG derivative order 2 is applied on the AH dataset whilst SG derivative order 1 is applied to the AST, DMM, and MGP datasets.

Once the ML model is successfully calibrated with 70% of the samples, the performance evaluation is carried out by testing with the remaining 30% of the samples in the respective dataset. The evaluation aims to assess the model accuracy in the sense of regression or classification problems. For regression, the ML model is evaluated using the coefficient of determination (R^2) and mean absolute error (MAE) expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_k - y_k)^2}{\sum_{i=1}^{n}(y_k - \bar{y}_k)^2} \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_k - y_k| \quad (4)$$

In Eqns. (3)-(4), $\hat{y}_k$ indicates the predicted output at sample k, y_k represents the observed value, and $\bar{y}_k$ is the mean of the observed values.

For classification, the evaluation is performed by looking at the classification accuracy (CA) and area under curve (AUC) values of the receiver operating characteristic curves. CA is defined as ratio of correctly predicted class label to the total number of the samples, i.e.:

https://doi.org/10.31436/iiumej.v25i1.2796

$$CA = \frac{number\ of\ correct\ prediction}{total\ number\ of\ samples} \quad (5)$$

Table 2: The spectra preprocessing and the ML hyperparameters setup.

Data-set	Type	Spectra Pre-processing	ML Hyperparameters Setting				
			XGB	RF	SVM	ANN	SEML
AH	Regression	Edge cutting, gaussian smoothing & SG derivative order 2	Number of estimators = 500 Learning rate = 0.25	Number of estimators = 800 Max depth = 85	C = 1000 Gamma = 0.01	Alpha = 0.01 Learning rate = 0.0001	Linear regression with elastic net regularization L1:L2=0.5:0.5
AST		Gaussian smoothing & SG derivative order 1	Number of estimators = 300 Learning rate = 0.25	Number of estimators = 800 Max depth = 85	C = 1 Gamma = 1	Alpha = 0.01 Learning rate = 0.0001	
DMM		Gaussian smoothing & SG derivative order 1	Number of estimators = 500 Learning rate = 0.25	Number of estimators = 800 Max depth = 85	C = 1000 Gamma = 0.01	Alpha = 0.01 Learning rate = 0.01	
MGP		Gaussian smoothing & SG derivative order 1	Number of estimators = 300 Learning rate = 0.25	Number of estimators = 500 Max depth = 60	C = 1000 Gamma = 0.01	Alpha = 0.01 Learning rate = 0.0001	
RA	Classification	Edge cutting, gaussian smoothing & SG derivative order 1	Trees: 50 Depth: 10	C: 10	Lr: 0.1 Trees: 50	Lr: 0.0001 Neurons: 50	Logistic regression with Ridge (L2) regularization
CF		Gaussian smoothing & SG derivative order 1	Trees: 50 Depth: 10	C: 10	Lr: 0.01 Trees: 50	Lr: 0.0001 Neurons: 50	
SB		Gaussian smoothing & SG derivative order 1	Trees: 50 Depth: 10	C: 10	Lr: 0.1 Trees: 200	Lr: 0.01 Neurons: 200	
SP		Edge cutting, gaussian smoothing & SG derivative order 1	Trees: 50 Depth: 30	C: 100	Lr: 0.01 Trees: 100	Lr: 0.0001 Neurons: 50	

https://doi.org/10.31436/iiumej.v25i1.2796

This CA metric can be obtained directly from the confusion matrix of classification datasets. Furthermore, the metrics in Eqns. (3) – (5) will be computed for the training (calibration) and testing data samples in each dataset. In addition, AUC can be generally defined as the measure of the ability of a classifier to distinguish between classes. The AUC=1 is for a perfect classifier while AUC=0.5 is for the worst classifier, as it only gives a random guess.

6. RESULTS AND ANALYSIS

This section discusses the results and analysis on the ML model performance to predict outputs for calibration and test samples in each dataset. The performance of the proposed SEML is compared to the individual ML model for each dataset.

Table 3 shows the performance metrics of the calibration model for the regression datasets as discussed in the previous section. Generally, all the ML algorithms used for calibration on the five regression datasets perform excellently in the training samples. Bold marks in the table indicate the highest performance in terms of R^2 value on each dataset. If two ML models proclude the same value, then both will be marked with bold.

Table 3: Evaluation results of the calibration model (regression datasets)

Dataset	ML model	Calibration		Test	
		R^2	MAE	R^2	MAE
AH	GB	≈1	0.134	0.877	8.919
	RF	0.999	0.596	0.820	11.007
	SVM	0.945	5.359	0.936	6.117
	ANN	0.939	5.954	0.933	6.397
	SEML	0.938	5.944	**0.940**	**6.076**
AST	GB	≈1	≈0	0.955	0.211
	RF	0.999	0.018	**0.963**	**0.200**
	SVM	0.951	0.206	0.961	0.207
	ANN	0.941	0.234	0.950	0.243
	SEML	0.959	0.184	**0.963**	**0.199**
DMM	GB	0.959	0.382	0.841	0.752
	RF	0.999	0.046	0.855	0.657
	SVM	0.863	0.678	0.864	0.687
	ANN	0.669	1.110	0.662	1.145
	SEML	0.889	0.608	**0.896**	**0.601**
MGP	GB	≈1	0.002	0.994	0.263
	RF	≈1	≈0	0.995	0.245
	SVM	0.999	0.102	**0.998**	**0.139**
	ANN	0.999	0.118	**0.998**	**0.162**
	SEML	0.999	0.115	**0.998**	**0.141**

They also well generalize the data in the respective regression tasks where R^2 of at least 0.662 is achieved in the test samples. The fact that only simple spectra pre-processing steps are applied as shown in Table 2 can be appreciated. For instance, most of the datasets only required gaussian smoothing and SG derivative order 1. In addition, the proposed SEML can improve, or at least perform on par with, the accuracy of individual base ML especially in the test samples.

Figure 6 shows the regression graph of observed values vs predicted values performed by SEML for test samples. This graph is displayed based on the results shown in Table 3. The model performs best for the DMM dataset and worst for the AH dataset as compared relative

https://doi.org/10.31436/iiumej.v25i1.2796

to other datasets. Since the DMM dataset has a bigger size (11362 samples) as compared to the MGP dataset (231 samples), the ML models should be able to learn sufficiently from the DMM data. However, the spectra pre-processing plays an important role contributing to the accuracy. This can be improved by implementing different pre-processing techniques, but this aspect is not the main focus of this paper and will not be explored further. Instead, a simple and relatively uniform preprocessing technique has been applied for all datasets.

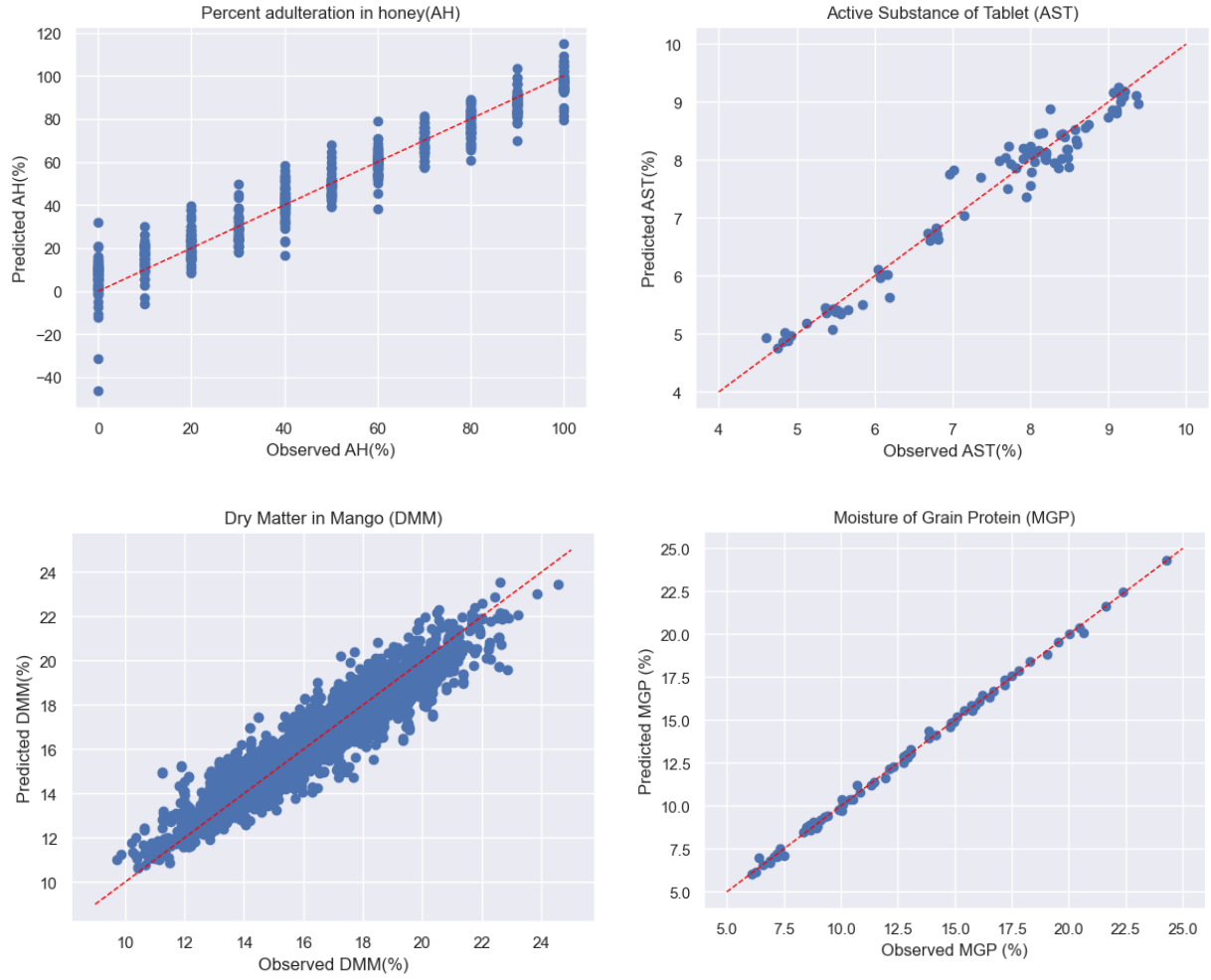

Fig. 6: Observed vs predicted values by SEML in test samples for regression datasets.

Table 4 shows the performance metrics of the calibration model for the classification datasets, as discussed in the previous section. Bold marks in the table indicate the highest performance in terms of CA value on each dataset. If two or more ML models procude the same value, then they will be marked with bold.

Generally, all the ML algorithms used for calibration on the selected four classification datasets perform excellently in the training samples that most of the predictions result in CA=1. They also well generalize the data in the respective classification tasks where CA of at least approximately 0.8 is achieved with most of the resulting CA being approximately 0.95 up to 1. The proposed SEML can achieve CA=1 for both training and testing samples in two datasets used, for example, CF and SP. In the classification datasets, the pre-processing only uses gaussian smoothing and an SG derivative of order 1. Edge cutting is only used for the spectra data collected by micro NIR spectrometer due to noises at around edges around 900-950 nm

https://doi.org/10.31436/iiumej.v25i1.2796

and 1650 nm – 1700 nm. In addition, the proposed SEML can slightly improve, or at least in par with, the accuracy of individual base ML in both training and testing samples. All four general models such as RF, SVM, GB and ANN can be considered as very competitive methods for some specific dataset to SEML. For example, SEML in RA dataset have similar CA with ANN and RF, while CF dataset SEML's CA have a similar result with ANN and SVM, etc.

Table 4: Evaluation results of the calibration model (classification datasets)

Dataset	ML model	Calibration		Test	
		CA	AUC	CA	AUC
RA	GB	1	1	0.946	0.925
	RF	1	1	**0.973**	**0.966**
	SVM	0.872	0.997	0.865	0.991
	ANN	1	1	**0.973**	**1**
	SEML	1	1	**0.973**	**0.978**
CF	GB	1	1	0.941	0.993
	RF	1	1	0.941	1
	SVM	1	1	**1**	**1**
	ANN	1	1	**1**	**1**
	SEML	1	1	**1**	**1**
SB	GB	1	1	0.976	0.993
	RF	1	1	0.966	0.993
	SVM	0.932	0.973	0.946	0.975
	ANN	0.968	0.993	0.969	0.993
	SEML	0.999	1	**0.980**	**0.994**
SP	GB	1	1	0.870	0.935
	RF	1	1	0.826	0.903
	SVM	1	1	**1**	**1**
	ANN	1	1	0.957	0.969
	SEML	1	1	**1**	**1**

Figure 7 shows the confusion matrices of the SEML for classification test samples. There is an interesting point here that the RA dataset is an imbalanced classification case. Despite the imbalanced data, all the ML models are still able to generalize the data very well especially for GB, RF, ANN and SEML. Furthermore, they also perform well for the CF and SP datasets despite their small number of samples.

7. DISCUSSION

Based on calibration results for both regression and classification tasks, in most cases, the Stacked Ensemble Machine Learning (SEML) approach performs better or at least comparably to the base learners. The observation is more evident in the regression case. It is important to focus on the robustness of SEML's generalization capability. The effectiveness of stacked ensemble models stems from the fact that different base learners tend to make different types of errors. Some base learners excel in capturing specific patterns or relationships in the data, while others perform better on different subsets or under varying circumstances. By combining the predictions of multiple base learners, the ensemble model achieves improved generalization and robustness. For instance, in regression case, in the AST and MGP datasets, Gradient Boosting (GB) performs exceptionally well on the training data but exhibits the worst performance on the test data compared to SEML, indicating a tendency of GB to overfit the data.

https://doi.org/10.31436/iiumej.v25i1.2796

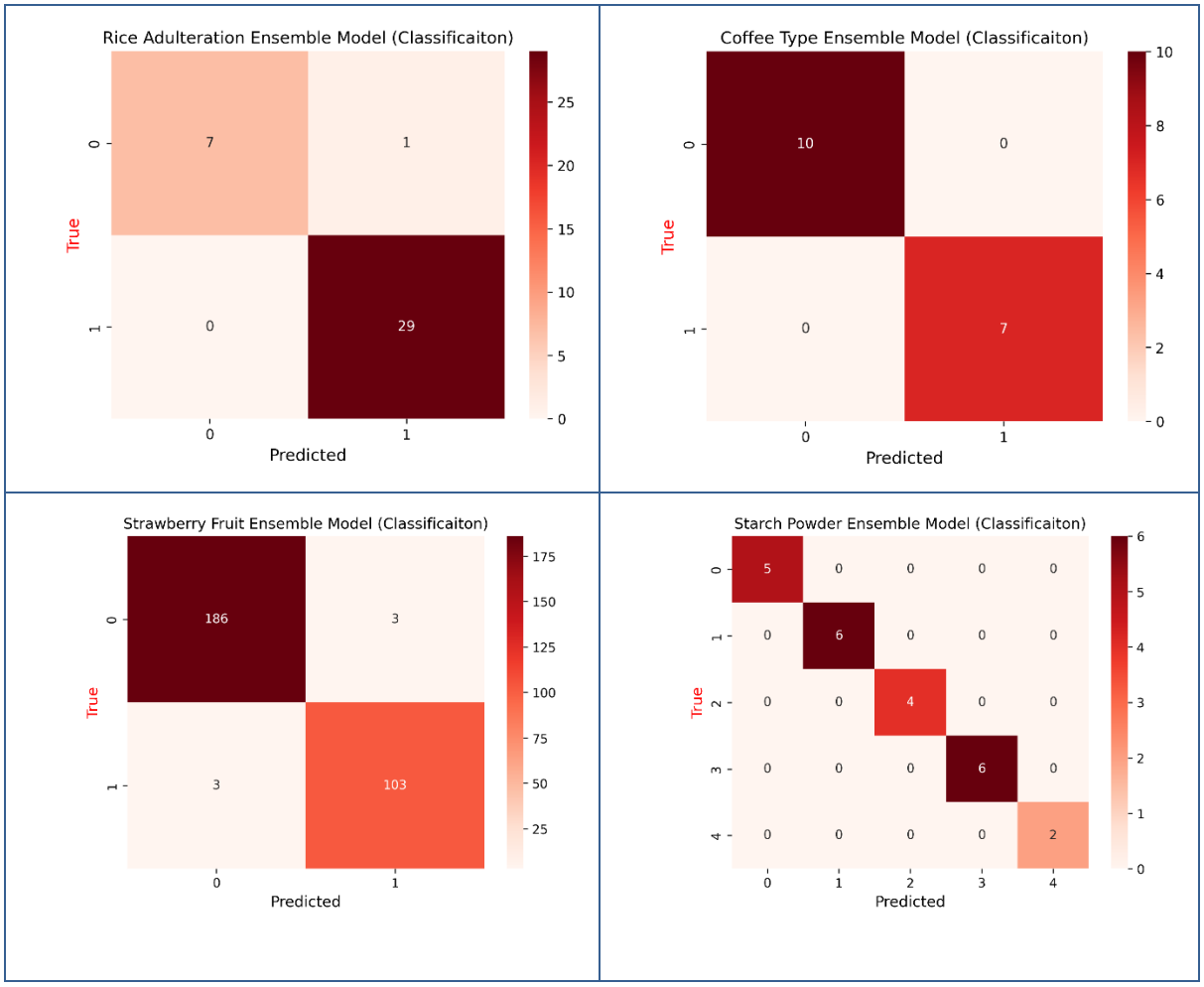

Fig. 7: Confusion matrix of SEML prediction in test samples for classification datasets.

In the classification scenario, the superiority of SEML becomes less evident, especially when considering the CF and SP datasets. This can be attributed to the fact that the base learner, Support Vector Machine (SVM), exhibits exceptional performance, accomplishing the task perfectly. Since SVM, and also ANN in the CF dataset, are already proficient in capturing the intricacies of these particular datasets, the additional benefits of utilizing SEML might be limited or negligible. Nonetheless, it is important to note that the potential advantages of SEML can still manifest in other datasets or when faced with more complex classification challenges where the base learner alone might not suffice.

8. DISCUSSION

Based on calibration results for both regression and classification tasks, in most cases, the Stacked Ensemble Machine Learning (SEML) approach performs better or at least comparably to the base learners. The observation is more evident in the regression case. It is important to focus on the robustness of SEML's generalization capability. The effectiveness of stacked ensemble models stems from the fact that different base learners tend to make different types of errors. Some base learners excel in capturing specific patterns or relationships in the data, while others perform better on different subsets or under varying circumstances. By combining the predictions of multiple base learners, the ensemble model achieves improved generalization and robustness. For instance, in regression case, in the AST and MGP datasets, Gradient

https://doi.org/10.31436/iiumej.v25i1.2796

Boosting (GB) performs exceptionally well on the training data but exhibits the worst performance on the test data compared to SEML, indicating a tendency of GB to overfit the data.

In the classification scenario, the superiority of SEML becomes less evident, especially when considering the CF and SP datasets. This can be attributed to the fact that the base learner, Support Vector Machine (SVM), exhibits exceptional performance, accomplishing the task perfectly. Since SVM, and also ANN in the CF dataset, are already proficient in capturing the intricacies of these particular datasets, the additional benefits of utilizing SEML might be limited or negligible. Nonetheless, it is important to note that the potential advantages of SEML can still manifest in other datasets or when faced with more complex classification challenges where the base learner alone might not suffice.

We will now present the average performance of each algorithm along with the corresponding metrics for both regression and classification tasks. The results can be found in Table 5. Bold marks in the table represent the highest performance in terms of the respective metrics. It is evident that SEML calibration yields better generalization capalility, i.e. always produces higher metrics on test data. In addition, it is worth to note that the closest performance with SEML are provided by GB and RF. This makes sense as GB and RF are a type of ensemble learning via boosting and bagging mechanism respectively, as discussed earlier.

Table 5: Average performance of each algorithm over regression datasets

ML model	Calibration		Test	
	R^2	MAE	R^2	MAE
GB	0.99	0.13	0.92	2.54
RF	**1.00**	**0.17**	0.91	3.03
SVM	0.94	1.59	0.94	1.79
ANN	0.89	1.85	0.89	1.99
SEML	0.95	1.71	**0.95**	**1.75**

Table 6: Average performance of each algorithm over classification datasets

ML model	Calibration		Test	
	CA	AUC	CA	AUC
GB	**1.00**	**1.00**	0.93	0.96
RF	**1.00**	**1.00**	0.93	0.97
SVM	0.95	0.99	0.95	0.99
ANN	0.99	1.00	0.97	0.99
SEML	**1.00**	**1.00**	**0.99**	**0.99**

9. CONCLUSION REMARKS

The results of chemometrics calibration for NIR spectroscopy data using stacked ensemble machine learning (SEML) have been presented. The prediction performance of the machine learning-based calibration model was evaluated and verified using eight (8) spectroscopy datasets representing both regression and classification cases. Despite employing a simple procedure for spectra data pre-processing, the machine learning methods, particularly GB, RF, SVM, and ANN as base learners, accurately predict both training and testing samples. Importantly, the proposed SEML, by combining the output of base learners, generally improves the accuracy of these individual base learners and provides better overall generalization, as confirmed through evaluation with test data.

The future direction of this research is to explore the potential applications of the proposed SEML to other datasets. The aim is to develop a robust calibration method that includes the

exploration of deep learning models, which could simplify the pre-processing of spectra. Deep learning holds potential for breakthroughs in this area due to its automatic feature extraction process, which is lacking in traditional machine learning approaches.

ACKNOWLEDGEMENT

This project is supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme, with code: FRGS/1/2020/TK0/UCSI/02/4.

REFERENCES

[1] Solihin MI, Shameem Y, Htut T, Ang CK, Hidayab M. (2019) Non-Invasive Blood Glucose Estimation using Handheld Near Infrared Device. Int. J. Recent Technol. Eng., 3: 16-19. doi: 10.35940/ijrte.C1004.1083S19.

[2] Chen CJ, Akowuah GA. (2023) Comparison of HPLC and ATR-FTIR Methods for the Determination of Rosmarinic Acid in Aqueous Leaf Extract of *Orthosiphon stamineus*. Nat. Prod. J., 13(1): 40-46. doi: 10.2174/2210315512666220429114935.

[3] B. A. Sabbagh, P. V. Kumar, Y. L. Chew, J. H. Chin, and G. A. Akowuah. (2022) Determination of metformin in fixed-dose combination tablets by ATR-FTIR spectroscopy. Chem. Data Collect., 13: 100868. doi: 10.1016/J.CDC.2022.100868.

[4] D. G. Abdullah Al-Sanabani, M. I. Solihin, L. P. Pui, W. Astuti, C. K. Ang, and L. W. Hong. (2019) Development of non-destructive mango assessment using Handheld Spectroscopy and Machine Learning Regression. Journal of Physics: Conference Series, 1367(1): 012030. doi: 10.1088/1742-6596/1367/1/012030.

[5] S. H. Tan, L. P. Pui, M. I. Solihin, K. S. Keat, W. H. Lim, and C. K. Ang. (2021) Physicochemical analysis and adulteration detection in Malaysia stingless bee honey using a handheld near-infrared spectrometer," J. Food Process. Preserv., 45(7): e15576. doi: 10.1111/JFPP.15576.

[6] K. T. Liew, L. P. Pui, and M. I. Solihin. (2020) Feasibility of fraud detection in rice using a handheld near-infrared spectroscopy. AIP Conference Proceedings, 2306(1): 020018. doi: 10.1063/5.0032679.

[7] P. S. Sampaio, A. Soares, A. Castanho, A. S. Almeida, J. Oliveira, and C. Brites. (2018) Optimization of rice amylose determination by NIR-spectroscopy using PLS chemometrics algorithms. Food Chem., 242: 196–204. doi: 10.1016/j.foodchem.2017.09.058.

[8] R. F. Kranenburg et al. (2020) Rapid and robust on-scene detection of cocaine in street samples using a handheld near-infrared spectrometer and machine learning algorithms. Drug Test. Anal., 12(10): 1404–1418. doi: 10.1002/DTA.2895.

[9] K. B. Beć and C. W. Huck. (2019) Breakthrough potential in near-infrared spectroscopy: Spectra simulation. A review of recent developments. Frontiers in Chemistry, 7(FEB). doi: 10.3389/fchem.2019.00048.

[10] H. P. Wang et al. (2022) Recent advances of chemometric calibration methods in modern spectroscopy: Algorithms, strategy, and related issues. TrAC Trends Anal. Chem., 153: 116648. doi: 10.1016/J.TRAC.2022.116648.

[11] H. A. Neto, W. L. F. Tavares, D. C. S. Z. Ribeiro, R. C. O. Alves, L. M. Fonseca, and S. V. A. Campos. (2019) On the utilization of deep and ensemble learning to detect milk adulteration. BioData Min., 12(1): 1–13. doi: 10.1186/s13040-019-0200-5.

[12] M. Y. Mohamed, M. I. Solihin, W. Astuti, C. K. Ang, and W. Zailah. (2019) Food powders classification using handheld Near-Infrared Spectroscopy and Support Vector Machine. J. Phys. Conf. Ser., 1367: 012029. doi:10.1088/1742-6596/1367/1/012029.

[13] D. Sing et al., (2021) Estimation of Andrographolides and Gradation of Andrographis paniculata Leaves Using Near Infrared Spectroscopy Together With Support Vector Machine. Front. Pharmacol., 12(May): 1–8. doi:10.3389/fphar.2021.629833.

[14] H. Chen, L. Xu, W. Ai, B. Lin, Q. Feng, and K. Cai. (2020) Kernel functions embedded in

support vector machine learning models for rapid water pollution assessment via near-infrared spectroscopy. Science of the Total Environment, 714: 136765. doi: 10.1016/j.scitotenv.2020.136765.

[15] A. P. M. Michel, A. E. Morrison, V. L. Preston, C. T. Marx, B. C. Colson, and H. K. White. (2020) Rapid Identification of Marine Plastic Debris via Spectroscopic Techniques and Machine Learning Classifiers. Environ. Sci. Technol., 54(17): 10630–10637. doi: 10.1021/acs.est.0c02099.

[16] I. M. Nolasco Perez, A. T. Badaró, S. Barbon, A. P. A. Barbon, M. A. R. Pollonio, and D. F. Barbin. (2018) Classification of Chicken Parts Using a Portable Near-Infrared (NIR) Spectrophotometer and Machine Learning. Appl. Spectrosc., 72(12): 1774–1780. doi: 10.1177/0003702818788878.

[17] Y. Wang, M. Li, R. Ji, M. Wang, and L. Zheng. (2020) Comparison of soil total nitrogen content prediction models based on Vis-NIR spectroscopy. Sensors (Switzerland), 20(24): 1–20. doi: 10.3390/s20247078.

[18] V. Woeng, L. Y. Lim, L. Abdul Kalam Saleena, M. I. Solihin, and L. P. Pui. (2022) Physicochemical properties and detection of glucose syrup adulterated Kelulut (Heterotrigona itama) honey using Near-Infrared spectroscopy. J. Food Process. Preserv., 46(7): e16686. doi: 10.1111/JFPP.16686.

[19] K. Nordhausen. (2022) Ensemble Methods: Foundations and Algorithms by Zhi-Hua Zhou. Int. Stat. Rev., 81(3): 470–470. doi: 10.1111/INSR.12042_10.

[20] H. Cao et al. (2022) Application of stacking ensemble learning model in quantitative analysis of biomaterial activity. Microchem. J., 183: 108075. doi: 10.1016/J.MICROC.2022.108075.

[21] M. Dyrby, S. B. Engelsen, L. Nørgaard, M. Bruhn, and L. Lundsberg-Nielsen. (2022) Chemometric Quantitation of the Active Substance (Containing C=N) in a Pharmaceutical Tablet Using Near-Infrared (NIR) Transmittance and NIR FT-Raman Spectra. Applied Spectroscopy, 56(5): 579-585. https://doi.org/10.1366/0003702021955358

[22] N. T. Anderson, K. B. Walsh, J. R. Flynn, and J. P. Walsh. (2021) Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content. II. Local PLS and nonlinear models. Postharvest Biol. Technol. 171: 111358. doi: 10.1016/J.POSTHARVBIO.2020.111358.

[23] "Data Sets - Eigenvector." [Online]. Available: https://eigenvector.com/resources/data-sets/. [Accessed: 28-Oct-2021].

[24] "Core Science Resources at QI." [Online]. Available: https://csr.quadram.ac.uk/. [Accessed: 29-Oct-2021].

[25] Holland. JK, Kemsley. EK, and Wilson. RH. (1998) Use of Fourier transform infrared spectroscopy and partial least squares regression for the detection of adulteration of strawberry purees. J. Sci. Food Agric., 76(2): 263–269. doi: 10.1002/(SICI)1097-0010(199802)76:2.

[26] U. Blazhko, V. Shapaval, V. Kovalev, and A. Kohler. (2021) Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra. Chemom. Intell. Lab. Syst., 215: 104367. doi: 10.1016/j.chemolab.2021.104367.

[27] D. Passos and P. Mishra. (2022) A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks. Chemom. Intell. Lab. Syst., 223: 104520 . doi: 10.1016/j.chemolab.2022.104520.

[28] D. S. Long, R. E. Engel, and M. C. Siemens. (2008) Measuring Grain Protein Concentration with In-line Near Infrared Reflectance Spectroscopy. Agron. J., 100(2): 247–252. doi: 10.2134/AGRONJ2007.0052.

[29] J. Acquarelli, T. van Laarhoven, J. Gerretzen, T. N. Tran, L. M. C. Buydens, and E. Marchiori. (2017) Convolutional neural networks for vibrational spectroscopic data analysis. Anal. Chim. Acta, 954: 22–31. doi: 10.1016/J.ACA.2016.12.010.

[30] H. Kew. (2021) A model for spectroscopic food sample analysis using data sonification. Int. J. Speech Technol., 24(4): 865–881. doi: 10.1007/s10772-020-09794-9.

[31] M. I. Solihin, Z. Zekui, C. K. Ang, F. Heltha, and M. Rizon. (2021) Machine Learning Calibration for Near Infrared Spectroscopy Data: A Visual Programming Approach. Lecture Notes in Electrical Engineering, 666: 577–590. doi: 10.1007/978-981-15-5281-6_40/COVER

[32] M. I. Solihin, Yanto, G. Hayder, and H. A. Q. Maarif. (2023) Landslide Susceptibility Mapping with Stacking Ensemble Machine Learning. Adv. Sci. Technol. Innov., 1: 35–40. doi: 10.1007/978-3-031-26580-8_7/COVER.

[33] T. Chen and C. Guestrin. (2016) XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1: 785–794. doi: 10.1145/2939672.2939785.

[34] Z. Cheng, Y. Yang, and H. Zhang. (2022) Interpretable ensemble machine-learning models for strength activity index prediction of iron ore tailings. Case Stud. Constr. Mater., 17: e01239. doi: 10.1016/J.CSCM.2022.E01239.

[35] K. P. Chan, M. I. Solihin, C. K. Ang, and L. P. Pui. (2022) Experimentation on Spectra Data Regression Using Dense Multilayer Neural Networks with Common Pre-processing. Lect. Notes Electr. Eng., 900: 97–112. doi: 10.1007/978-981-19-2095-0_10/COVER.