European Journal of Agronomy 167 (2025) 127558

Contents lists available at ScienceDirect

European Journal of Agronomy

journal homepage: www.elsevier.com/locate/eja

Non-destructive and on-site estimation of grape total soluble solids by field spectroscopy and stack ensemble learning

Hongyi Lyu [a,*], Miles Grafton [a,*], Thiagarajah Ramilan [a], Matthew Irwin [a], Eduardo Sandoval [b]

[a] School of Agriculture and Environment, Massey University, Palmerston North 4410, New Zealand
[b] Massey Agri-Food (MAF) Digital Lab., Massey University, Palmerston North 4410, New Zealand

ARTICLE INFO

Keywords:
Machine learning
Grape berries
Field spectroscopy
Spectroradiometer
Total soluble solids

ABSTRACT

Accurately estimating the total soluble solids (TSS) of berries with a non-destructive method is crucial for wine grape growers if wine quality improvements are to be made. At present, the methods employed with the best statistical results are implemented under stable lab conditions, using spectroscopic analysis in the visible-near infrared (VNIR) region. This study explores using field spectroscopy to estimate the TSS of berries directly in the vineyard. A portable visible-near infrared-shortwave infrared (VNIR-SWIR) spectroradiometer measured the reflectance data of grape berries in the 350–2500 nm spectral region. A large in-field multi-season spectral database ($n = 1830$) over two years (2023–2024) from three 'Pinot Noir' commercial vineyards were selected to develop spectral-region specific (VNIR, SWIR or VNIR-SWIR) machine learning models. Different machine learning modeling pipelines were built using data collected from 2023 and validated using data from 2024 to predict grape TSS based on in-field spectral databases. Subsequently, the performance of using stack ensemble learning (ES) to predict grape TSS was evaluated and compared with three commonly used methods: K-nearest neighbors (KNN), random forest regression (RFR), and support vector regression (SVR). The result on the independent test set showed that, the ES model based on MSC+SG+ 1D spectral data, in the VNIR-SWIR region provided the highest prediction accuracy for grape TSS value, with a coefficient of determinations (R^2) of 0.815, root mean square error (RMSE) of 1.131 °Brix, and a ratio of performance to deviation (RPD) of 2.236, with a Lin's concordance correlation coefficient (CCC) of 0.897. This study demonstrated the potential of using an ES model to assess the grape TSS rapidly and non-destructively from field spectroscopy data.

1. Introduction

Grapevine (*Vitis* spp.) is one of the key perennial crops in New Zealand, due to its commercial use in wine production. Accurately estimating the quality of berries is crucial for industrialized wine-growing systems, as the quality of wine is strongly influenced by the grape quality at the time of picking (Chen et al., 2015). Grape quality is traditionally monitored by TSS in New Zealand. TSS in grape pulp can be measured using lab-based destructive techniques such as utilizing a refractometer to measure their °Brix value. The traditional measurement method only provides an average representation of the TSS in a specific area of the vineyard, while previous studies have shown that TSS varies spatially in a single block (Baluja et al., 2013). Winemakers regard the uniformity of fruit berries as a key aspect of wine quality. When uniform berries are utilized, wine can be produced with the consistent flavor and aroma (Bramley, 2022). To obtain the spatial-temporal variation of TSS in fruit throughout the vineyard, repeated measurements must be taken at various locations and times during ripening. This requires many destructive measurements over each vineyard. However, this is not desirable for commercial vineyards as it is labor intensive and expensive. Therefore, in the commercial wine industry, there is a significant demand for non-destructive methods that can offer fast and dependable results for assessing grape TSS (Dambergs et al., 2015).

Many studies have explored the possibility of using spectroscopic analysis in assessing various grape quality parameters or varieties non-destructively (Larraín et al., 2008; González-Caballero et al., 2010, 2012; Urraca et al., 2016; dos Santos Costa et al., 2019; Benelli et al., 2021; Kalopesa et al., 2023). Most of these studies utilized spectroscopic measurement to predict TSS in wine grape berries under controlled laboratory conditions and obtained high prediction accuracy (dos Santos Costa et al., 2019). However, the application of spectroscopy in vineyard conditions faces challenges due to unpredictable environmental factors

* Corresponding authors.
E-mail addresses: hlyu@massey.ac.nz (H. Lyu), m.grafton@massey.ac.nz (M. Grafton).

https://doi.org/10.1016/j.eja.2025.127558
Received 26 June 2024; Received in revised form 2 February 2025; Accepted 14 February 2025

such as light, temperature, and sample location. Larraín et al., (2008) and Guidetti et al., (2010) demonstrated the potential of using field spectroscopy for monitoring grape quality parameters directly in the vineyard. However, the spectral range of reflectance data used in these studies is limited in the VNIR region (400–1000 nm). The utilization of spectral reflectance data, particularly in the SWIR region (1000–2500 nm), has gained considerable interest in monitoring grape quality parameters, as indicated by González-Caballero et al., (2012). Their study revealed that the SWIR spectra, particularly in the wavelength range of 1600–2400 nm, can be an effective tool in classifying the ripening stage of wine grapes. In addition, Urraca et al., (2016) achieved moderate prediction accuracy, with RMSE of 1.68 °Brix for TSS using a spectra range of 1600–2400 nm. In studies conducted in the VNIR-SWIR region, Kalopesa et al., (2023) achieved good prediction accuracy, with a R^2 of 0.87 (RMSE of 1.76 °Brix) for TSS using a convolutional neural networks model. To the best of our knowledge, there is currently no definitive research indicating whether the spectral data in VNIR or SWIR, or a combination of both (VNIR-SWIR), is preferable for predicting TSS. Costs present a crucial consideration for industrial applications, with VNIR spectroscopy being generally cheaper when compared to SWIR. Therefore, it is crucial to evaluate the effect of different spectral regions on predicting TSS.

Furthermore, machine learning techniques have been employed to leverage the rich information present in spectral reflectance data, enabling the identification of relationships between reflectance and fruit quality parameters (Kalopesa et al., 2023). Nevertheless, the high dimensionality and multicollinearity issues in spectral reflectance data can lead to overfitting of machine learning models, reducing their accuracy and general usefulness. It is essential to perform effective data pre-processing and feature selection, to mitigate these issues and enhance the performance, of machine learning algorithms when working with spectroscopic data (Minas et al., 2023; Ma et al., 2024). Dimensionality reduction methods are commonly employed in studies to reduce storage space, computation time, and remove redundant features in high-dimensional data, in order to mitigate multicollinearity issues (Silva and Melo-Pinto, 2021, 2023). The dimensionality reduction methods can be broadly classified into two categories: linear techniques, such as Principal Component Analysis (PCA), and non-linear techniques, such as t-Distributed Stochastic Neighbor Embedding, Kernel PCA and autoencoders. Silva and Melo-Pinto, (2021) applied various dimensionality reduction techniques in a machine learning model to predict grape sugar content and found PCA outperformed other non-linear techniques. Their approach depended on single predictive models, SVR and are subject to overfitting with limited training data. Over the past decade, various ensemble techniques that use a collection of base learners to improve predictive performance were proposed (Dong et al., 2024). Common ensemble strategies include bagging, boosting, and stacking. Bagging generates base learners in parallel by creating training subsets through bootstrap sampling. Boosting trains a sequence of base models by capitalizing on their dependences (Breiman, 1996). Stacking, unlike bagging and boosting, typically employs heterogeneous learners and leverages their differences to improve accuracy (Wolpert, 1992). By ensuring diversity among the base learners, complementary information is provided, thus the risk of incorrect results is reduced. To the best of our knowledge, the application of a stacking ensemble strategy in grape quality prediction has not been observed, though it has been shown in successful applications in predicting crop yield in various studies on other crops (Ji et al., 2023).

This study utilizes field spectroscopy data to identify the TSS value of wine grapes in a non-destructive way. The specific objectives of this research are:

- To develop a non-destructive method to estimate wine grape TSS based on field spectroscopy.
- To compare the performance of stack ensemble learning model with individual machine learning model.
- To evaluate the effect of different spectral regions (VNIR, SWIR, and VNIR-SWIR) on model performance.
- To quantify the effects of different data pre-processing methods on model estimations.
- To validate model transferability in different seasons.

2. Methodology

2.1. Study site

The study was conducted during the period between veraison and harvest seasons of 2023 and 2024 (early February to late March) at three commercial vineyards: Hua Nui (HN) (both years 2023–24), Pencarrow (PN) (in year 2023), and Clouston (CN) (in year 2023). The vines were $>$ 15 years old 'Pinot Noir' (with UCD 4 clones) and trained with vertical shoot positioning. In the HN vineyard, vine spacing is 1.7 m and row spacing is 2.2 m. In the PN vineyard, vine spacing is 1.8 m with the same row spacing of 2.2 m. In the CN vineyard, vine spacing is 1.6 m, also with a row spacing of 2.2 m. The soils in the vineyards are mostly clay and silt loams, which are known to have moderate soil water holding capacity. The vineyard managers were wholly responsible for cultivation practices uniformly across the sample areas of each vineyard.

In 2023 and 2024, a total of 1843 grape berries were collected to estimate TSS. In 2023, seven field visits were carried out in three vineyards (HN, PN, and CN). On each measurement date, several healthy vines were selected across the vineyards based on grid sampling strategy (Fig. 1). Whereas in 2024, four field visits were carried out in HN vineyard. In 2024, a sample of 77 vines was selected from the HN vineyard, which had been sampled in 2023 (Figure S1). During the initial field visits, berry samples were collected from only 14 vines, while in subsequent field visits, samples were collected from 77 vines each time. In 2023 and 2024, each sampling vine provided three berries selected from the top, middle and bottom position of a single bunch (Fig. 2a).

2.2. Acquisition of spectral reflectance data and pre-processing

In-situ spectral acquisitions were performed using the ASD FieldSpec 4 Hi-Res NG Spectroradiometer (Malvern Panalytical Ltd., Malvern, UK), which is capable of measuring the VNIR-SWIR spectrum (350–2500 nm). The spectroradiometer employed a contact probe for controlled illumination (Fig. 2b). During the measurement, a white panel ceramic referencing tile (Malvern Panalytical Ltd., Malvern, UK) was utilized for calibration, and as a reference spectrum, after each set of three measurements on a given vine. The ratio of the reflected optical radiation from the sample to that of the reference panel was used to calculate reflectance values. Reflectance spectra data of each individual berry were obtained by averaging five measurements. The spectral data were subsequently interpolated to achieve a spectral resolution of 1 nm, resulting in a dataset comprising 2151 bands ranging from 350 to 2500 nm. To eliminate signal noise, reflectance data outside the spectral regions below 450 nm and above 2200 nm were excluded. The processed data were then exported as ASCII text files for each spectral measurement using ViewSpec Pro 6.2 software (Analytical Spectral Devices, Inc., Boulder, CO, USA).

To enhance the subsequent modeling phase and eliminate physical effects present in the spectra, the raw reflectance data (RW) were subjected to various spectral transformations. Specifically, the first derivative (1D), Savitzky–Golay filtering with polynomial = 3, window size = 11, and derivative = 0 (SG), absorbance (Abs = Log (1/RW)), and Multiplicative Scatter Correction (MSC) combined with SG for 1D (MSC+SG+1D) were applied. These transformations aimed to mitigate effects such as baseline shifts, slope variations, which could affect the modeling process (Rinnan et al., 2009). Derivative pre-processing procedures were performed using the "prospectr" package in R statistical software (R Core Team, version 4.2.2). A derivative gap of 3 was

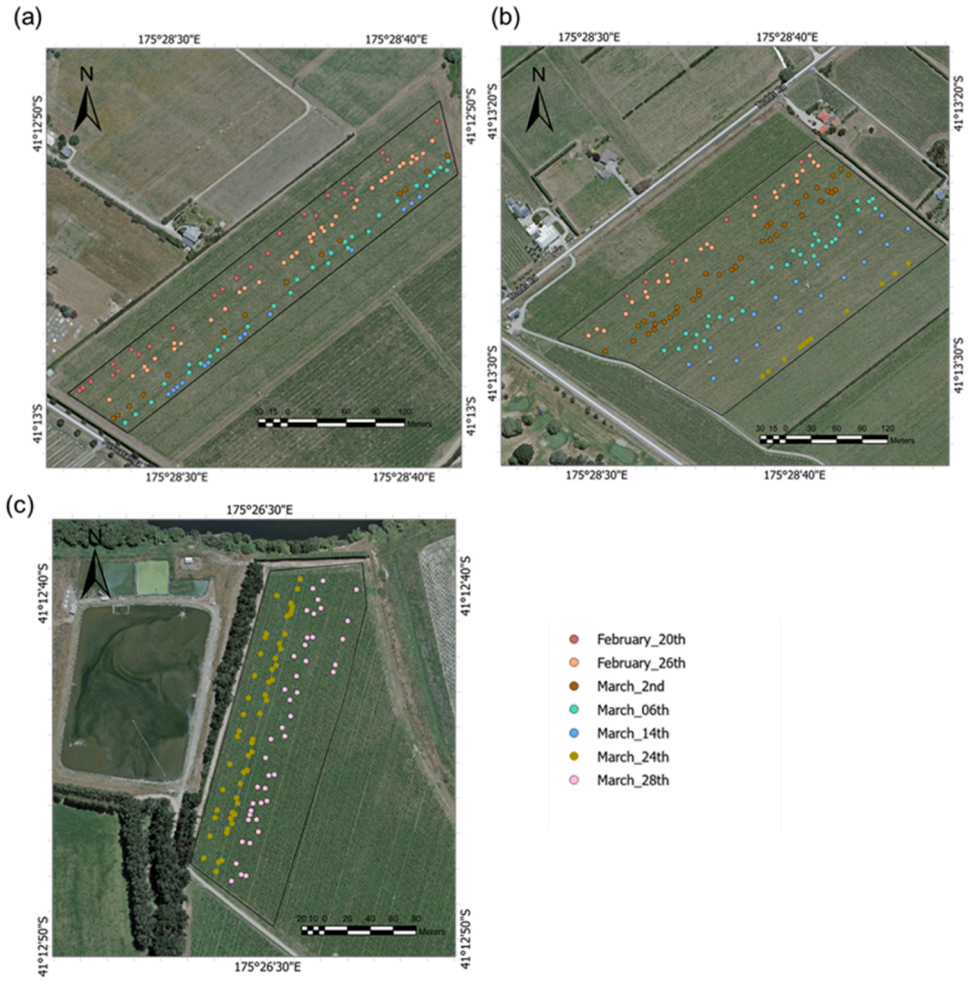

Fig. 1. Sampling location in Hua Nui (a), Pencarrow (b), and Clouston (c) in 2023 harvesting season.

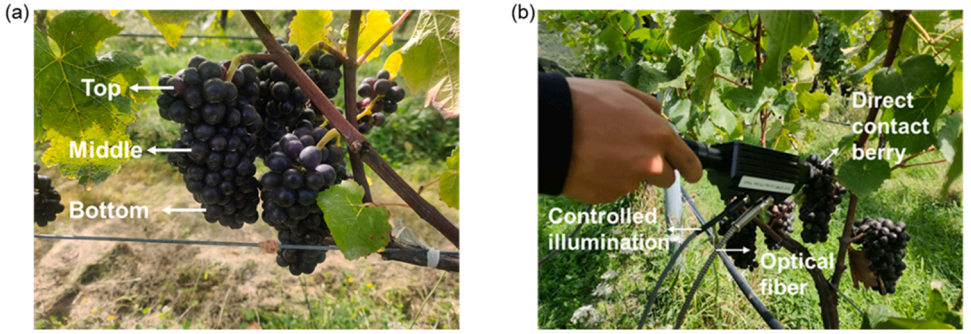

Fig. 2. The location of sampled berries in a grape cluster (a), Collection of the grape berry using the ASD FieldSpec 4 Hi-Res NG Spectroradiometer (b).

employed during the computation of the derivatives (Wei et al., 2021, Lyu et al., 2023).

2.3. TSS content measurement

Following the acquisition of spectral reflectance data, the measurement of TSS in grape berries was carried out on the same grape berries which were used for acquiring spectral reflectance data. TSS, expressed in °Brix, was measured by using a portable digital refractometer (PAL-ALPHA Digital Refractometer, ATAGO CO., LTD, Tokyo, Japan). The refractometer had a measuring range of 0–85 °Brix with an accuracy of 0.20 °Brix. Prior to each measurement, the refractometer prism underwent a thorough rinse with deionized water, followed by drying using absorbent cleaning paper. Calibration was performed by adjusting the instrument to read zero degrees of °Brix using a drop of deionized water at around 25 °C. For analysis, each grape berry selected was carefully cut and squeezed to extract the juice, which was then placed onto the transparent prism surface of the refractometer.

2.4. Dimensionality reduction method

The dimensionality reduction method improves the prediction performance of machine learning models for spectral reflectance data by removing redundant features in a dataset (Silva and Melo-Pinto, 2021).

In this study, PCA was chosen for the dimensionality reduction method for the five feature groups (RW, 1D, SG, Abs and MSC+SG+1D). The PCA has been widely utilized in dimensionality reduction of spectral reflectance data and proved to outperform other methods (Silva et al., 2018; Silva and Melo-Pinto, 2021).

PCA enabled the transformation of high-dimensional data into a lower-dimensional space, while preserving as much of the original variation as possible. The transformation results in the creation of "principal components (PCs)" which were linear combinations of the original features. In general cases, the optimal number of PCs to retain for analysis is determined by the proportion of variance explained (selecting the number of PCs which accounted for the majority of the variance in the population). However, there isn't any arbitrary optimum method for selecting the number of PCs to retain. Thus, in this study we set the PCs as a hyperparameter in different machine learning modeling pipelines. Consequently, the model was tested for each dataset using a range of one to 20 PCs, and the ideal number of PCs was selected based on cross-validation. We choose 20 PCs as the upper limit since the 20 PCs explain 99.99 % of the variability in the original spectral data. Additionally, retaining a higher number of PCs for analysis leads to an increase in computational costs.

2.5. Data-driven modeling

Before the application of the machine learning model, anomalous samples detection was performed based on 1.5 times interquartile range (IQR) rule and PCA. The outlier in the spectral data was detected by the Mahalanobis distance of the first and second principal components in the PCA. The outlier in the °Brix data was detected by the limits of 1.5 IQR. The anomalous samples were excluded when the limits of 1.5 IQR and PCA were crossed simultaneously. In this case 1830 samples were selected for further analysis. To estimate TSS from spectral reflectance data (RW, 1D, SG, Abs, and MSC+SG+1D), the following machine learning algorithms were applied: KNN, RFR and SVR. KNN regression is an instance-based learning method that predicts the numerical target based on distance functions (e.g., Euclidean and Manhattan). KNN regression as a non-parametric method, predicting the response by calculating the average of the responses from the KNN in the training set. Delgadillo-Duran et al., (2022) successfully used KNN to predict soil properties from spectral data. Additionally, RFR and SVR are widely used in spectroscopic analysis to predict grape quality parameters (Tsakiridis et al., 2023). RFR is an ensemble learning algorithm that generates a series of regression trees by using different bagged samples. The final predicted value of RFR is obtained by averaging the predictions generated by all the individual regression trees in the ensemble. SVR is an extension of the support vector machines specifically designed for regression problems. It finds the best hyperplane to maximize the margin between the predicted values and the observed values, making it useful for predicting continuous numerical values.

In addition, an ES model was used in order to improve the prediction performance. The ES model is a common hierarchical ensemble learning strategy, that typically comprises two levels: level 1, consisting of basic learners; and level 2, a meta-learner. The basic learners encompass multiple machine learning algorithms, and the outputs from these basic models were collected and combined to generate meta-features. The basic learners in this study included the KNN, RFR and SVR. The meta-learner aggregates the outputs from the basic learners to produce the final prediction. This study selected the highest-performing basic learner as the meta-learner for the ensemble. Specifically, this study evaluated the accuracy of each basic learner in predicting TSS, and the model with the best accuracy was used as the meta-learner. This approach leverages the strengths of each model while allowing the ensemble to focus on the most accurate predictions. Fig. 3 illustrate this ensemble model structure. For processing all machine learning models, the mlr3verse package in R software version 4.2.2 was used. Table S1 show the tuned hyperparameters and their ranges for each model.

2.6. Model performance evaluation

The measured spectrum (450–2200 nm) was analyzed as a whole and also separately for its two subsections: the VNIR (450–1000 nm) and SWIR (1001–2200 nm) regions. The three spectral regions (VNIR, SWIR, and VNIR-SWIR) were processed using five spectral preprocessing methods (RW, 1D, SG, Abs, and MSC+SG+1D). A PCA dimensionality reduction technique was applied to these pre-processed spectral vectors, and the resulting principal components (PCs) were then used in four machine learning models (KNN, RF, SVR, and ES) for predicting TSS. A total of 120 modelling pipelines were developed for TSS modelling (Table S2). In this study, the train-test split method was implemented according to the year of data collection. Data collected in 2023 ($n = 1099$) were designated for model training, while data collected in 2024 ($n = 731$) were reserved for model performance evaluation. This approach employs an independent dataset for validation, reducing the risk of overfitting, and evaluating the generalization performance of the

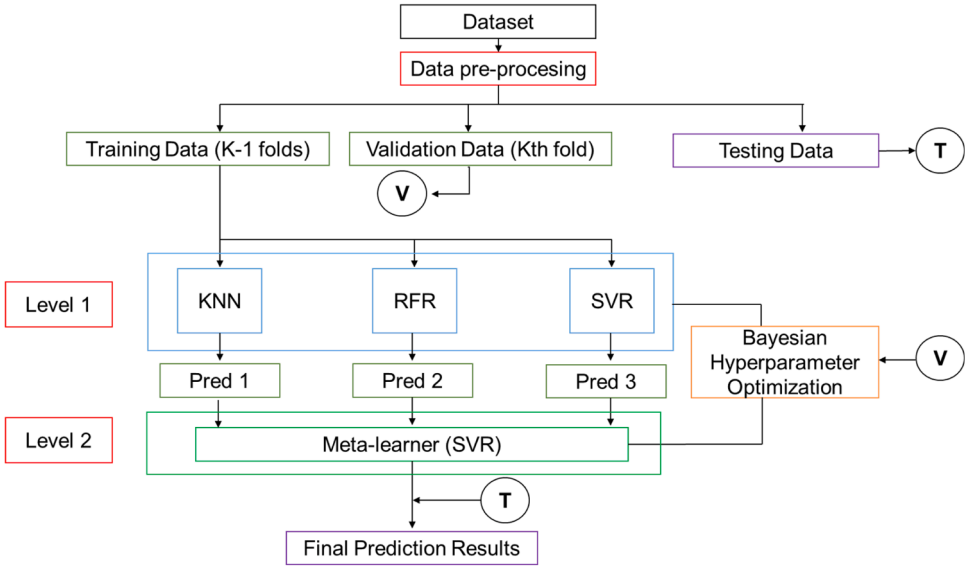

Fig. 3. The workflow of the ES model for TSS estimation (Level 1 consists of three base learners: KNN, RFR, and SVR; Level 2 includes a single meta-learner: SVR).

model. Additionally, a 10-fold cross validation strategy was used to tune the hyperparameters in each training set of all the tested machine learning models. The combination of hyperparameters contributing to the models with the lowest RMSE values, were considered as optimized. In order to evaluate the model performance, the values of RMSE, ratio of performance to deviation (RPD), R^2, and Lin's concordance correlation coefficient (CCC) were calculated by applying the trained models with optimized hyperparameters to the test set. The equations of the indices are shown in Equations (3) – (5).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

$$RPD = \frac{SD(y)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}} \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

$$CCC = \frac{2\rho\sigma_{y_i}\sigma_{\hat{y}_i}}{\sigma_{y_i}^2 - \sigma_{\hat{y}_i}^2 + (\mu_{y_i} - \mu_{\hat{y}_i})^2}$$

where n is the number of samples used to fit the model, y_i is the ground truth value of the ith sample, $\bar{y}$ is the mean response value, $\hat{y}_i$ is the model estimated value of the ith sample.

3. Results

3.1. Statistical analysis of measured TSS and spectral reflectance data

The period from veraison to harvest is recognized as the most crucial phase for monitoring the quality of grape berries (González-Caballero et al., 2012). After outlier detection, throughout this period, a total of 1830 berries were collected, and the summary statistics of the training and testing datasets on berry TSS values is presented in Table 1. In this study, a wide range of TSS values was found. Berry TSS ranged from 8.2 to 30.9 °Brix. The wide range of the dataset made it easier to build a robust calibration model. Table 1 show high TSS variability (SD = 2.53) for the samples collected from HN vineyard in 2024 compared to the samples of the other vineyards; whilst low TSS variability (SD = 1.59) for the samples from CN vineyard.

Fig. 4 show the average spectra of the grape samples collected from different vineyards in different seasons, covering the spectral range of 450–2200 nm. The small reflectance between 500 and 700 nm was related to pigments such as anthocyanins, chlorophyll, and carotenoids present in fruit (Ncama et al., 2017). The absorption at around 970 nm was related to carbohydrate and water. González-Caballero et al., (2010) found that the 940–980 nm region as the most relevant for TSS measurement. In addition, sugar-related absorption bands were also found at around 1200 nm (Williams and Norris, 1987). However, most of the short-wave infrared (SWIR, 1000–2500 nm) regions appeared to provide limited information in terms of absorption bands. The weak absorption at around 1450 nm and 1940 nm was related to water in fruit

Table 1
Descriptive statistics of training and testing datasets on total soluble solids (where n = number of samples, min = minimum, max = maximum, SD = standard deviation).

Year	Datasets	Vineyard	n	min	mean	max	SD
2023	Train	HN	352	8.3	17	23.4	2.02
		PN	466	8.2	16.6	21.2	2.11
		CN	281	16.1	20.4	24.8	1.59
2024	Test	HN	731	9.7	19.7	30.9	2.53

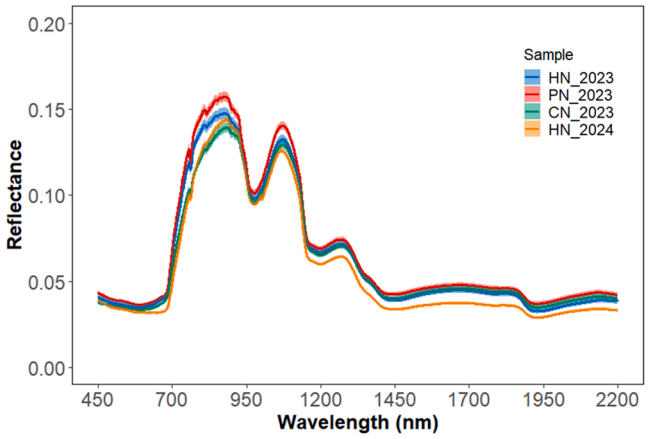

Fig. 4. The mean spectral values of 'Pinot Noir' grape samples from different vineyards in different seasons (Different background colors of the spectral values represents the 95 % confidence interval).

(Omar, 2013).

PCA was conducted to reduce the redundant information in the spectral data (RW, 1D, SG, Abs and MSC+SG+1D). After PCA, both the raw and pre-processed spectral data were transformed into principal components. Fig. 5 show the result of principal component analysis of raw spectral data. According to the PCA analysis, the first two principal components accounted for 71.4 % and 21.4 % of the variance, respectively (Fig. 5a, b), indicating that they capture the primary data structure. However, when tuning the model with PCs, using only 1 or 2 PCs did not yield good results. Therefore, this study set the PC range from one to 20, as 20 PCs explained most of the variance. Beyond 20 PCs, there was no improvement in model performance, but computational time increased. The data collected from HN vineyard during 2024 show the most concentrated distribution around the origin, while data collected from PN vineyard during 2023 display more scattered distribution with greater variance along both PC1 and PC2 axes Fig. 5c illustrated the loadings plot for the first five principal components. There are a large number of peaks throughout the spectrum from 700 to 1000 nm, that are usually associated with sugar content.

3.2. Prediction accuracy based on different machine learning models

All the machine learning models (KNN, RFR, SVR and ES) were trained using different pre-processing methods and tested on independent datasets. A total of 120 modeling pipelines were tested, and the results were shown in Table S2. The best results of different machine learning models based on different spectral pre-processing and dimensionality reduction methods are shown in Table 2. Among the four machine learning models, the MSC+SG+ 1D spectral preprocessing method consistently demonstrated the best performance when compared with other spectra-preprocessing methods. The model performance criteria showed that the ES model (R^2 = 0.815, RMSE = 1.131 °Brix, CCC = 0.897 and RPD = 2.236) outperformed KNN (R^2 = 0.669, RMSE = 1.454 °Brix, CCC = 0.801 and RPD = 1.74), RF (R^2 = 0.705, RMSE = 1.373 °Brix, CCC = 0.843 and RPD = 1.842), and SVM (R^2 = 0.783, RMSE = 1.178 °Brix, CCC = 0.893 and RPD = 2.148) models on the independent test set (Table 2). The scatter plot of predicted and measured TSS value on the independent test set for each best performance machine learning models are shown in Fig. 6.

The result of using different spectral region data for predicting TSS values shows the TSS estimates apparently favor the VNIR region for KNN and RF models as compared to the SWIR or VNIR-SWIR spectral regions (Fig. 7). However, SVM and ES models showed higher accuracy in the VNIR-SWIR spectral region when predicting TSS (Fig. 7). The result of RMSE from different machine learning models and pre-

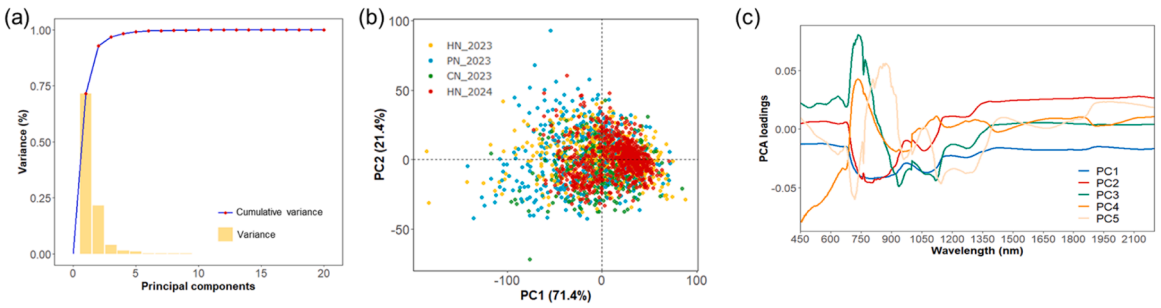

Fig. 5. The result of principal component analysis of raw spectral data. Cumulative variance of the principal components (a), Scored images of the first two principal components (b), Loadings plot for the first five principal components (c).

Table 2
Model performance criteria for predicting TSS from field spectroscopy data on the independent test set includes coefficients of determination (R^2), root mean square error (RMSE), ratio of performance to deviation (RPD), and Lin's concordance correlation coefficient (CCC).

Model	Spectral region	Pre-processing	R^2	RMSE	CCC	RPD
KNN	VNIR	MSC+SG+ 1D+PCA	0.669	1.454	0.801	1.74
RF	VNIR	MSC+SG+ 1D	0.705	1.373	0.843	1.842
SVM	VNIR-SWIR	MSC+SG+ 1D	0.783	1.178	0.893	2.148
ES	VNIR-SWIR	MSC+SG+ 1D	0.815	1.131	0.897	2.236

processing methods showed a consistent result where the prediction performance using SWIR region alone are lower than those using VNIR and VNIR-SWIR (Fig. 7).

Fig. 7 also illustrate the influence of PCA on prediction accuracy across the different machine learning models (KNN, RF, SVM, and ES) and spectral pre-processing methods in terms of RMSE (°Brix). The addition of PCA as a dimensionality reduction method generally appears to reduce the RMSE values, particularly in the cases where PCA is applied with specific model pipelines. Since the SWIR region does not yield good prediction results, this study focused on the VNIR and VNIR-SWIR regions. For the KNN model, the application of PCA in the VNIR region with the Abs and MSC+SG+ 1D preprocessing methods improves predictive accuracy. In the VNIR-SWIR region, PCA consistently improves performance across all spectral preprocessing methods, which underscores PCA's utility in this broader spectral range for KNN.

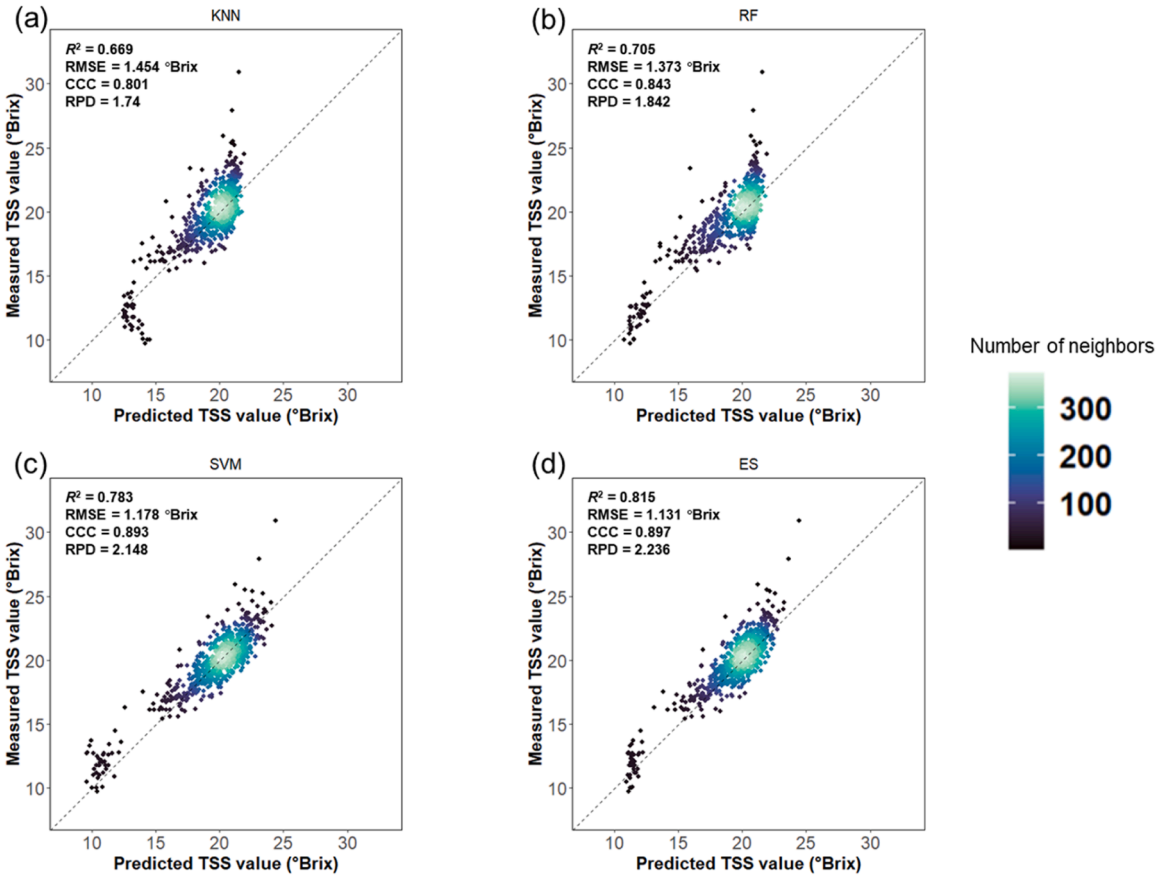

Fig. 6. The 1:1 line relationship between predicted and measured TSS values for the independent test set by KNN based on MSC+SG+ 1D+PCA spectral data in the VNIR region (a), RF based on MSC+SG+ 1D spectral data in the VNIR region (b), SVM based on MSC+SG+ 1D spectral data in the VNIR-SWIR region (c), ES based on MSC+SG+ 1D spectral data in the VNIR-SWIR region (d).

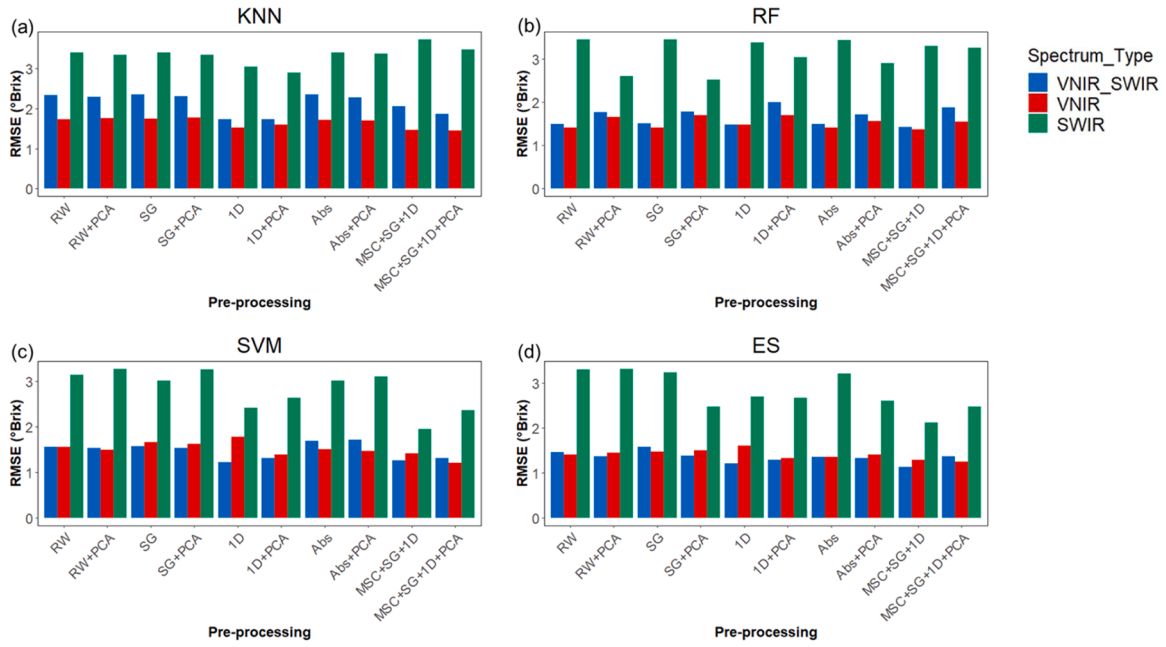

Fig. 7. The TSS prediction accuracy (RMSE) of the four machine learning models (KNN, RF, SVM and ES) using the grape berry spectral features in the VNIR, SWIR and VNIR-SWIR spectral regions for independent test set.

However, for the RF model, PCA does not enhance prediction accuracy in either the VNIR or VNIR-SWIR regions. In the SVR model, PCA enhances all spectral preprocessing methods in the VNIR region. In the VNIR-SWIR region, PCA improves the results for the RW and SG preprocessing methods. For the ES model, PCA enhances the 1D and MSC+SG+ 1D preprocessing methods in the VNIR region, and in the VNIR-SWIR region, PCA improves the results for RW, SG, and Abs preprocessing. In summary, PCA tends to improve models that are more sensitive to high-dimensional data (such as KNN and SVR), particularly in the VNIR-SWIR region.

4. Discussion

Traditional destructive measurement only measures a few berries in a single block to represent the whole vineyard's berry quality. The present study explores a novel non-destructive approach to estimate the TSS of wine grapes when berries are grown in grape clusters through a portable VNIR-SWIR spectroradiometer (350–2500 nm). Different machine learning models and pre-processing methods were applied to predict TSS from in-field spectral data. To assess the generalization of the machine learning models, different seasonal data sets were used as an independent test set.

4.1. The stack ensemble learning model performance on TSS estimation

Several studies have utilized machine learning models in combination with spectral data to predict grape berry quality (Fernandes et al., 2015; Silva et al., 2018). For example, SVR a common non-linear machine learning models has been applied to predict grape berry TSS and pH (Silva et al., 2018). In addition, Gomes et al., (2017) have demonstrated the application of neural networks as an individual prediction model for predicting grape TSS. It is worth noting that these studies relied on an individual basic learner for constructing predictive models, related to grape berry quality. In order to improve the prediction accuracy, this study used a stacking-based ensemble learning (ES) approach, which combines three individual learners to construct the TSS prediction. KNN, RFR and SVR were utilized for constructing ES, and SVR was selected as the meta-learner for this ensemble approach. The result of using the ES model with different pre-processing methods and spectral region data outperformed other individual machine learning models in independent test sets (Table 2, S2, Fig. 7). Compared with individual machine learning models, the ES model can reduce bias and variance in the final prediction by capturing intricate patterns that individual models cannot capture alone. Several successful examples of use of the ES approach have also been observed in crop yield prediction (Ji et al., 2023). This study evaluates 120 machine learning pipelines for predicting grape TSS value (Table S2). The result shows the best prediction performance was achieved by applying an ES model to MSC+SG+ 1D data in the VNIR-SWIR spectral region (R^2 = 0.815, RMSE = 1.131 °Brix, CCC = 0.897 and RPD = 2.236), which was better than the findings of Gutiérrez et al., (2019) where the RMSE was 1.274 °Brix. This study showing the success of using the ES approach to predict the berry quality.

4.2. Influence of spectral region on model performance

Silva and Melo-Pinto, (2021) suggested that the VNIR spectral region is optimal for predicting grape TSS because of its cheaper acquisition Most studies measured the grape berry reflectance in the VNIR (400–1000 nm) spectral region alone to predict grape TSS values (Guidetti et al., 2010; Silva et al., 2018; Tsakiridis et al., 2023; Benelli et al., 2021). However, dos Santos Costa et al., (2019) and Kalopesa et al., (2023) measured reflectance data in the VNIR-SWIR spectral region. The prediction results from dos Santos Costa et al., (2019) and Kalopesa et al., (2023) were slightly higher than that of other studies using VNIR spectral regions to predict grape quality. This study explores the influence of different spectral regions (VNIR, SWIR, VNIR-SWIR) on model performance. Different machine learning models show the best predictions in different spectral regions (Table 2, S2, Fig. 7). SVM and ES models in this study have shown slightly higher performance within the VNIR-SWIR spectral region. However, the KNN and RF models showed lower accuracy within the VNIR-SWIR spectral region to that of the VNIR region. For the best performance model (ES model), a wider range of spectral regions are required, rather than using spectral data only in VNIR region, which improves the model predictions. However, compared with VNIR, the use of VNIR-SWIR spectroscopy for grape TSS prediction does not significantly improve the prediction performance, so in practical applications, wine grape growers are currently

recommended to use VNIR equipment rather than VNIR-SWIR because of its present lower cost and good prediction ability.

4.3. Influence of pre-processing method on model performance

In order to solve the high dimensionality and multicollinearity issues within the spectral reflectance data, this study used different pre-processing methods to reduce the dimensionality of the data to an appropriate level. The PCA approach was widely employed as one of the most commonly used dimensionality reduction methods. It offers an acceptable performance while requiring moderate computational resource (Silva and Melo-Pinto, 2021). The efficiency of using transformed variables from the PCA method for predicting TSS was evaluated. In this study, we set the number of principal components (PCs) as a hyperparameter, ranging from one to 20, with 20 PCs chosen to explain most of the variability in the original spectra. Using over 20 PCs in the model pipeline increased computational costs, it did not result in improved accuracy. The ES model performance in the independent test set showed that using the PCA method to pre-process spectral data can improve the model performance among most model pipelines in the VNIR-SWIR spectral region (Fig. 7), which was consistent with previous studies (Silva et al., 2018; Silva and Melo-Pinto, 2021). Despite PCA not enhancing prediction accuracy across all models, its application offers a substantial reduction in model training time. By condensing high-dimensional spectral data into fewer components, PCA lowers computational demands, which can be particularly beneficial in scenarios with limited processing resources. Future research should explore using additional dimensionality reduction techniques such as independent component analysis (ICA), t-distributed stochastic neighbour embedding (t-SNE), and autoencoders in model pipeline.

Previous studies showed that different spectral pre-processing can improve the prediction accuracy when predicting grape berry TSS, grape leaf nutrient status, and grape water status (Kalopesa et al., 2023; Lyu et al., 2023; Wei et al., 2021). In this study, when using KNN, RFR, SVR and ES algorithms. the MSC+SG+ 1D spectral preprocessing method consistently demonstrated the best performance compared with other spectra-preprocessing methods. However, other preprocessing methods do not improve the performance of the model in most cases (Fig. 7). One possible reason for the poor performance of transformation variables is that since the spectral pre-processing method used in this study did not consider the scatter correction method. In future studies, more combinations of spectral pre-processing and dimensionality reduction method should be selected to investigate the impact of pre-processing methods on predicting berry quality.

4.4. Model generalization

Most studies restricted their study area to one vineyard or one season (dos Santos Costa et al., 2019; Gutiérrez et al., 2019; Benelli et al., 2021; Kalopesa et al., 2023). The test and train data set in these studies often come from the same environmental conditions, thus high model performance can be expected. However, these models are less likely to extend to an independent data. Only a few studies have explored the transferability of models, seeking to create a model from one vineyard and utilize it on different vineyards or harvest season (Urraca et al., 2016; Silva et al., 2018). Independent validation stands out as the most reliable approach for evaluating a model's generalization ability. The independent validation dataset was collected in a different harvest season than that of the training dataset, reflecting seasonal variations in climate conditions. Thus, this independent dataset was used to evaluate four machine learning models' generalization ability. The result showed a satisfactory range of the prediction performance and model generalization ability of four machine learning models (Table 2). Comparing the results in the previous studies using in-lab spectral data from different vintages in the testing phase, we found that Silva et al., (2018) obtained a R^2 of 0.837, and a RMSE of 2.443 °Brix, for the determination of TSS, when using different varieties and vintages as test sets. Even though spectral data used in this study were collected from unstable environmental conditions (Fig. 2), the model performance ($R^2 = 0.815$, RMSE = 1.131 °Brix, CCC = 0.897 and RPD = 2.236) is in accord with Silva et al., (2018) when using the ES model. This indicates robustness and reliability of the ES mode in processing field spectral data. This study presents a solution for using field spectroscopy to replace conventional techniques for predicting grape TSS value. The results confirm the suitability of predicting 'Pinot Noir' TSS values. Non-destructive and rapid estimate of grape TSS values aids vineyard managers to determine harvest strategies such as selective harvesting and mapping spatial variability of grape TSS. It is noticeable that the independent dataset was collected in a different harvest season, the vineyard used for independent testing is one of the three vineyards included in the training dataset. Future studies should test the current ES model across vineyards in a wider range of growing environments and grape varieties to enhance its robustness and applicability.

5. Conclusion

Estimating the grape berry TSS value during harvest is important for vineyard management to ensure better berry quality. This study has explored a large database collected in different vineyards and harvest seasons to establish a simple, relatively less computational costly, and non-destructive method for predicting grape berry TSS directly in-field using different machine learning modeling pipelines. The results revealed that the ES model based on MSC+SG+ 1D spectral data in the VNIR-SWIR region was suitable to estimate 'Pinot Noir' TSS values ($R^2 = 0.815$, RMSE = 1.131 °Brix, CCC = 0.897 and RPD = 2.236). The ES model outperformed individual models, demonstrating its potential to combine strengths for improved performance. The evaluation of spectral regions showed that the VNIR-SWIR region provided the most reliable results. The model's transferability across different seasons suggests promise, though further validation is needed with more grape varieties. The developed methodology in this study has demonstrated the possibility for a rapid and non-destructive measurement of TSS in wine grapes based on field spectroscopy.

Funding

This research received no external funding.

CRediT authorship contribution statement

Grafton Miles: Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization. **Lyu Hongyi:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sandoval Eduardo:** Software, Resources, Data curation. **Irwin Matthew:** Supervision, Software, Resources, Data curation. **Ramilan Thiagarajah:** Visualization, Validation, Supervision, Investigation, Funding acquisition, Data curation.

Declaration of Competing Interest

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

Acknowledgements

The authors acknowledge the assistance of Mr. Guy McMaster, manager and winemaker of Palliser Estate Ltd.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.eja.2025.127558.

Data availability

Data will be made available on request.

References

Baluja, J., Tardaguila, J., Ayestaran, B., Diago, M.P., 2013. Spatial variability of grape composition in a Tempranillo (Vitis vinifera L.) vineyard over a 3-year survey. Precis. Agric. 14, 40–58.

Benelli, A., Cevoli, C., Ragni, L., Fabbri, A., 2021. In-field and non-destructive monitoring of grapes maturity by hyperspectral imaging. Biosyst. Eng. 207, 59–67.

Bramley, R.G.V., 2022. Precision Viticulture: managing vineyard variability for improved quality outcomes. In *Managing wine quality*. Elsevier, pp. 541–586.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Chen, S., Zhang, F., Ning, J., Liu, X., Zhang, Z., Yang, S., 2015. Predicting the anthocyanin content of wine grapes by NIR hyperspectral imaging. Food Chem. 172, 788–793.

Dambergs, R., Gishen, M., Cozzolino, D., 2015. A review of the state of the art, limitations, and perspectives of infrared spectroscopy for the analysis of wine grapes, must, and grapevine tissue. Appl. Spectrosc. Rev. 50 (3), 261–278.

Delgadillo-Duran, D.A., Vargas-García, C.A., Varón-Ramírez, V.M., Calderón, F., Montenegro, A.C., Reyes-Herrera, P.H., 2022. Vis–NIR spectroscopy and machine learning methods to diagnose chemical properties in Colombian sugarcane soils. Geoderma Reg. 31, e00588.

Dong, R., Miao, Y., Wang, X., Kusnierek, K., 2024. An active canopy sensor-based in-season nitrogen recommendation strategy for maize to balance grain yield and lodging risk. Eur. J. Agron. 155, 127120.

Fernandes, A.M., Franco, C., Mendes-Ferreira, A., Mendes-Faia, A., da Costa, P.L., Melo-Pinto, P., 2015. Brix, pH and anthocyanin content determination in whole Port wine grape berries by hyperspectral imaging and neural networks. Comput. Electron. Agric. 115, 88–96.

Gomes, V., Fernandes, A., Faia, A., Melo-Pinto, P., 2017. Comparison of different approaches for the prediction of sugar content in new vintages of whole Port wine grape berries using hyperspectral imaging. Comput. Electron. Agric. 140, 244–254.

González-Caballero, V., Sánchez, M.-T., López, M.-I., Pérez-Marín, D., 2010. First steps towards the development of a non-destructive technique for the quality control of wine grapes during on-vine ripening and on arrival at the winery. J. Food Eng. 101 (2), 158–165.

González-Caballero, V., Sánchez, M.-T., Fernández-Novales, J., López, M.-I., Pérez-Marín, D., 2012. On-vine monitoring of grape ripening using near-infrared spectroscopy. Food Anal. Methods 5, 1377–1385.

Guidetti, R., Beghi, R., Bodria, L., 2010. Evaluation of grape quality parameters by a simple Vis/NIR system. Trans. ASABE 53 (2), 477–484.

Gutiérrez, S., Tardáguila, J., Fernández-Novales, J., Diago, M.P., 2019. On-the-go hyperspectral imaging for the in-field estimation of grape berry soluble solids and anthocyanin concentration. Aust. J. Grape Wine Res. 25 (1), 127–133.

Ji, Y., Liu, R., Xiao, Y., Cui, Y., Chen, Z., Zong, X., Yang, T., 2023. Faba bean above-ground biomass and bean yield estimation based on consumer-grade unmanned aerial vehicle RGB images and ensemble learning. Precis. Agric. 1–22.

Kalopesa, E., Karyotis, K., Tziolas, N., Tsakiridis, N., Samarinas, N., Zalidis, G., 2023. Estimation of sugar content in wine grapes via in situ VNIR–SWIR point spectroscopy using explainable artificial intelligence techniques. Sensors 23 (3), 1065.

Larraín, M., Guesalaga, A.R., Agosín, E., 2008. A multipurpose portable instrument for determining ripeness in wine grapes using NIR spectroscopy. IEEE Trans. Instrum. Meas. 57 (2), 294–302.

Lyu, H., Grafton, M., Ramilan, T., Irwin, M., Sandoval, E., 2023. Assessing the leaf blade nutrient status of pinot noir using hyperspectral reflectance and machine learning models. Remote Sens. 15 (6), 1497.

Ma, F., Wang, M., Yan, N., Adnan, M., Jiang, F., Hu, Q., He, G., Shen, Y., Wan, Y., Yang, Y., 2024. A fast and efficient phenotyping method to estimate sugarcane stalk bending properties using near-infrared spectroscopy. Eur. J. Agron. 154, 127107.

Minas, I.S., Anthony, B.M., Pieper, J.R., Sterle, D.G., 2023. Large-scale and accurate non-destructive visual to near infrared spectroscopy-based assessment of the effect of rootstock on peach fruit internal quality. Eur. J. Agron. 143, 126706.

Ncama, K., Opara, U.L., Tesfay, S.Z., Fawole, O.A., Magwaza, L.S., 2017. Application of Vis/NIR spectroscopy for predicting sweetness and flavour parameters of 'Valencia'orange (Citrus sinensis) and 'Star Ruby'grapefruit (Citrus x paradisi Macfad). J. Food Eng. 193, 86–94.

Omar, A.F., 2013. Quantifying water-sucrose solutions through NIR spectral absorbance linearisation and gradient shift. Journal of Optics 42, 189–193.

Rinnan, Å., Van Den Berg, F., Engelsen, S.B., 2009. Review of the most common pre-processing techniques for near-infrared spectra. TrAC Trends Anal. Chem. 28 (10), 1201–1222.

dos Santos Costa, D., Mesa, N.F.O., Freire, M.S., Ramos, R.P., Mederos, B.J.T., 2019. Development of predictive models for quality and maturation stage attributes of wine grapes using vis-nir reflectance spectroscopy. Postharvest Biol. Technol. 150, 166–178.

Silva, R., Melo-Pinto, P., 2021. A review of different dimensionality reduction methods for the prediction of sugar content from hyperspectral images of wine grape berries. Appl. Soft Comput. 113, 107889.

Silva, R., Melo-Pinto, P., 2023. T-SNE: a study on reducing the dimensionality of hyperspectral data for the regression problem of estimating oenological parameters. Artif. Intell. Agric. 7, 58–68.

Silva, R., Gomes, V., Mendes-Faia, A., Melo-Pinto, P., 2018. Using support vector regression and hyperspectral imaging for the prediction of oenological parameters on different vintages and varieties of wine grape berries. Remote Sens. 10 (2), 312.

Tsakiridis, N.L., Samarinas, N., Kokkas, S., Kalopesa, E., Tziolas, N.V., Zalidis, G.C., 2023. In situ grape ripeness estimation via hyperspectral imaging and deep autoencoders. Comput. Electron. Agric. 212, 108098.

Urraca, R., Sanz-Garcia, A., Tardaguila, J., Diago, M.P., 2016. Estimation of total soluble solids in grape berries using a hand-held NIR spectrometer under field conditions. J. Sci. Food Agric. 96 (9), 3007–3016.

Wei, H.-E., Grafton, M., Bretherton, M., Irwin, M., Sandoval, E., 2021. Evaluation of point hyperspectral reflectance and multivariate regression models for grapevine water status estimation. Article 16. Remote Sens. 13 (16). https://doi.org/10.3390/rs13163198.

Williams, P., Norris, K., 1987. Near-infrared technology in the agricultural and food industries. American Association of Cereal Chemists, Inc.

Wolpert, D.H., 1992. Stacked generalization. Neural Netw. 5 (2), 241–259.