Peer Reviewed Article openaccess

Estimation of strawberry firmness using hyperspectral imaging: a comparison of regression models

Binu Melit Devassy* and Sony George

Department of Computer Science, Norwegian University of Science and Technology, Gjøvik 2802, Norway

Contacts
Binu Devassy: binu.m.devassy@ntnu.no
https://orcid.org/0000-0003-1860-9749

Sony George: sony.george@ntnu.no
https://orcid.org/0000-0001-8436-3164

Firmness is one of the most important quality measures of strawberries, and is related to other aspects of the fruit, such as flavour, ripeness and internal characteristics. The most popular method for measuring firmness is puncturing with a penetrometer, which is destructive and time-consuming. In the present study, we make an attempt to predict the firmness of strawberries in a fast, non-destructive and non-contact way using hyperspectral imaging (HSI) and data analysis with various regression techniques. The primary goal of this research is to investigate and compare the firmness prediction capability of seven prominent regression techniques. We have performed HSI data acquisition of 150 strawberries and optimised seven regression models using the spectral information to predict strawberry firmness. These models are linear, ridge, lasso, k-neighbours, random forest, support vector and partial least square regression. The results show that HSI data with regression models has the potential to predict firmness in a rapid, non-destructive manner. Out of these seven regression models, the k-neighbours regression model outperformed all other methods with a standard error of prediction of 0.14, which is better than that of the state-of-the-art results.

Keywords: hyperspectral imaging, non-destructive firmness measurement, strawberry firmness, regression models

Introduction

The quality inspection of fruits and vegetables is critical in a food value chain.[1] Producers, suppliers and consumers assess the quality based on several attributes, primarily by visual inspection and automated systems. Firmness is one of the leading quality indicators linked with other characteristics such as taste, ripeness levels etc. of the fruit or vegetable.[2] Firmness is a complicated feature determined by various attributes, such as the fruit's inner structure and composition.[3] Puncturing with a penetrometer is the most common approach for determining firmness. Estimating the fruit's firmness helps to determine the fruit's maturity and is a direct pointer to the harvesting time and shelf life. This research investigates hyperspectral imaging (HSI) applications to predict strawberries' firmness in a rapid, non-destructive manner. The attractiveness of strawberries is influenced by their

Correspondence
Binu Devassy (binu.m.devassy@ntnu.no)

Received: 2 May 2021
Revised: 18 June 2021
Accepted: 24 June 2021
Publication: 30 June 2021
doi: 10.1255/jsi.2021.a3
ISSN: 2040-4565

Citation
B.M. Devassy and S. George, "Estimation of strawberry firmness using hyperspectral imaging: a comparison of regression models", *J. Spectral Imaging* **10,** a3 (2021). https://doi.org/10.1255/jsi.2021.a3

ripeness, which is directly linked with the fruit's firmness.[4] Strawberry is a fruit with delicate skin and hence is very susceptible to injuries. It is crucial to estimate the firmness of the strawberry and sort the fruit accordingly.

Different destructive and non-destructive methods are available to measure the fruits' firmness, though no standardised methods have been established.[5] Destructive techniques measure the firmness by mechanically applying pressure, and in most cases, this causes some degree of damage to the fruit. Such physical disruptions also lead to undesirable metabolic chaos and biochemical changes of the fruit. Penetration methods using a penetrometer also provide information on firmness by estimating the needle's depth, which is inversely proportional to the firmness. However, this method also damages the fruit and is not suitable for large-scale measurements. Measurement techniques such as acoustic meters[6] and texture analysers[7] have been mainly limited to large-scale and online measurements. Optical techniques offer several advantages over the previously listed measurement methods, such as speed and non-contact nature. HSI, a technology originally developed for remote sensing, has recently gained much attention in close-range applications such as food inspection[8] and classification,[9,10] medical imaging,[11,12] forensics,[13,14] cultural heritage[15,16] etc. HSI has received wide acceptance in food analysis as it offers the possibility to simultaneously record both spectral and spatial information. As shown in Figure 1, HSI data can be viewed as a data cube[17] that consists of different band images as layers and is often referred to as an HSI data cube. It has spatial information along the X- and Y-axes and spectral information along the Z-axis. Fast and accurate spectral measurements provide the advantage of fruit quality evaluation in real-time, irrespective of the fruit's size and shape. Firmness estimation and prediction on different fruits such as banana, blueberries etc. using HSI have been reported.[18–20]

Researchers have attempted to predict strawberry firmness from HSI data using various algorithms in the last few decades. Nagata *et al.*[3] used HSI in the visible region to produce prediction models for firmness and soluble solids content (SSC) in strawberries using multiple linear regression.[21] Tallada *et al.*[22] used NIR HSI to predict firmness in strawberries using partial least squares (PLS) regression.[23] Sánchez *et al.*[24] applied modified PLS and local algorithms for firmness prediction. Liu *et al.*[25] used several computational models for predicting the firmness and total soluble solids (TSS) of intact strawberry fruit, including PLS, support vector machine (SVM) and back-propagation neural network (BPNN) from multispectral data. Recently Mancini *et al.*[26] used partial least squares regression (PLSR) to predict strawberry firmness. Even though these researches achieved a decent prediction capacity, many other popular regression methods such as ridge[27] and lasso (least absolute shrinkage and selection operator)[28] have never been tested for this purpose.

This research's primary goal is to check the applicability of other regression methods for better prediction capability than the state-of-the-art techniques and quantitatively compare the prediction results. This paper presents the HSI data in the visible near infrared (VNIR) region to predict strawberries' firmness using several regression models. We used some of the most popular regression models suitable for training on smaller datasets such as k-neighbours,[29] random forest,[30] along with the traditional ridge, lasso, linear, support vector[31] and PLS[23]

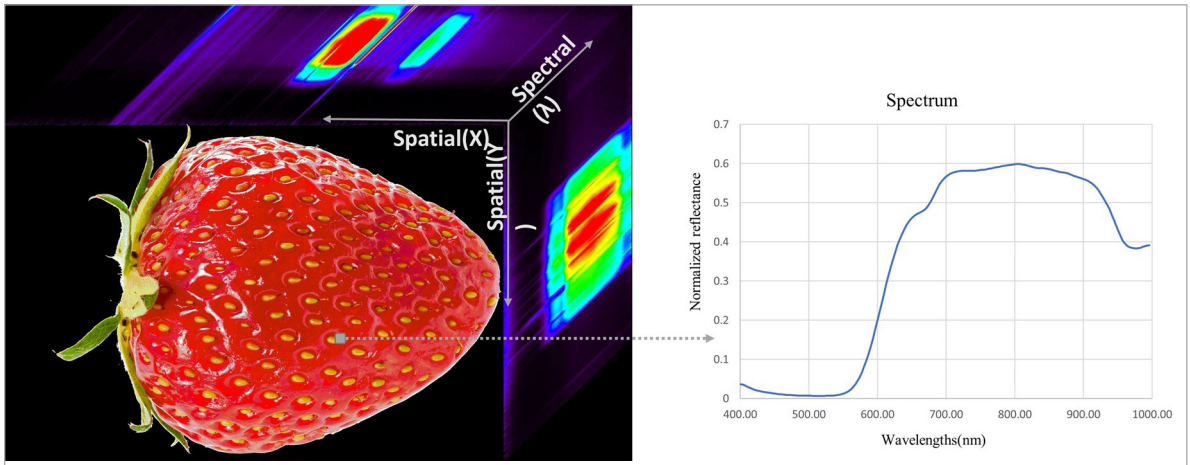


Figure 1. Hyperspectral data cube representation where the XY-axis indicates the spatial plane while the Z-axis represents the spectral dimension.

regressions. Five different parameters are used in the comparative study to assess the predictive skills of these methods. The best regression will be established based on the essential wavelengths to generate a lighter model and we used successive projections algorithm (SPA)[32] to select these wavelengths.

Materials and methods

Samples

Strawberries used in this study were purchased from a local farm in Norway. A total of 150 strawberries were selected as samples. Strawberries were carefully chosen by excluding fruits having any visible defects or bruises. All the strawberries were cleaned to eliminate any external contaminants, and the water drops from the fruits were swept away before the HSI acquisition and penetration measurement. The strawberries were all grown under the same conditions and were at the ripening or over-ripened stage. The fruits were stored in recommended storage conditions (4° C) and kept in the lab at a controlled room temperature for an hour before HSI acquisition. The strawberries are processed batch by batch; each batch contains 8 to 12 strawberries based on their size. Each batch is carefully taken out of the refrigerator, imaged and firmness measured before proceeding to the next batch.

Hyperspectral imaging

Hyperspectral imaging for this study was conducted using a camera (HySpex-VNIR-1800, developed by Norsk Elektro Optikk AS). The camera has a spectral sensitivity from 400 nm to 1000 nm and spectral sampling of 3.18 nm, which provides 186 spectral bands. This VNIR push broom scanner records 1800 pixels across the field of view spanning approximately 10 cm for the lens used, which has a 30 m focusing distance, 1 cm depth of focus and a polariser to avoid specular reflection. The camera was placed at right angles to a moving translator stage where the fruit samples were placed. Two halogen light sources were used to illuminate the scene with 45°:0° geometry to the camera. Details of the setup are presented in Figure 2. The image acquisition resulted in a hyperspectral data cube with spatial (X and Y) and spectral (Z) directions. Here, the X-axis size was 1800 pixels, the size of the Y-axis depends on the size and number of strawberries used in a single scan, and the size of the Z-axis was 186. A reference target with known reflectance (Contrast Multi-Step Target by Spectralon®) values was present in the scene for converting the radiance to the reflectance at the post-processing stage.

Firmness attribute measurements

The firmness of each strawberry was measured immediately after spectral measurement using a digital fruit penetrometer (Turoni, Italy). The firmness was measured using a puncture test at the fruit's equatorial side using an 8 mm diameter cylindrical probe and is expressed as Newton (N). The probe was attached to the stand with a handle to control the penetration speed. The maximum force (N) detected during the puncture test was recorded as the firmness of that particular fruit. The data contains the firmness of 150 strawberries with a minimum of 0.4 N and a maximum of 3.0 N, with an average of 1.1 N.

Regression models

The study of dependency is known as regression, and it is used in many research projects.[33] In this work, we denote the firmness value as variable *Y* and the predictor variable as *X*; the variable *X* represents the strawberry spectrum with values X_1, X_2, X_3,..., X_n. Here *n* is the number of

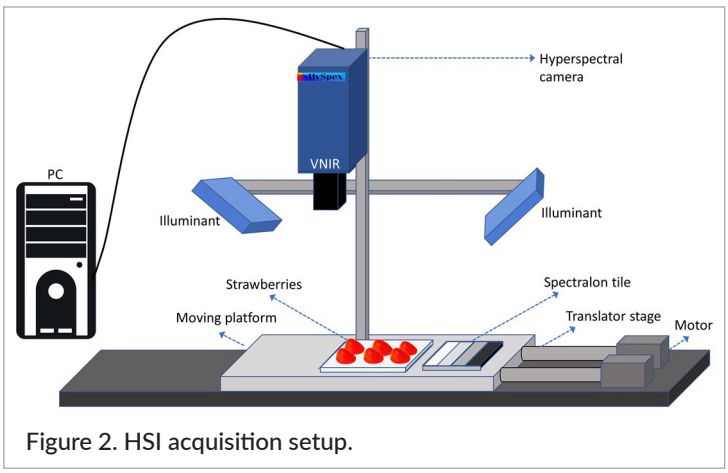


Figure 2. HSI acquisition setup.

wavelengths present in the spectrum and is 186 in the present case. Then, by using a regression model, the relationship between the spectrum (X) and firmness (Y) can be approximated using Equation 1.[34]

$$Y = f(X_1, X_2, X_3, ..., X_n) + \varepsilon \quad (1)$$

where ε represents the discrepancy in the approximation and is considered as a random error, and f represents the regression model. This work uses seven different regression models, and they are described in detail in the following sections.

Linear regression

Consider X to be a single explanatory variable; for a given set of X and Y observations, the relationship between a dependent variable Y and an independent variable X can be estimated using simple linear regression. Simple linear regression can be extended to include more than one explanatory component and is then called multiple linear regression.[21] Since we presume that the response variable is directly connected to a linear combination of the explanatory variables in both cases, we continue to use the expression "linear". The equation for multiple linear regression is similar to single linear regression, but it contains more terms than Equation 1,[21] and it is given in Equation 2.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_n X_{in} \quad (2)$$

where β_0 is the y-intercept, β_n is the slope coefficient for each explanatory variable, n is the number of bands and ε represents the prediction error. The linear regression tries to optimise the β values such that it minimises the cost function. Here we are using mean squared error (MSE) as the linear regression model's cost function, which is represented in Equation 3.

$$MSE = \frac{1}{p}\left[\sum_{i=1}^{p}\left(Y_i - \hat{Y}_i\right)^2\right] = \frac{1}{p}\left\{\sum_{i=1}^{p}\left[Y_i - (\beta_0 + \sum_{j=1}^{n}\beta_j X_{ij})\right]^2\right\} \quad (3)$$

where Y_i represents the expected (measured) firmness, $\hat{Y}_i$ represents the predicted firmness using linear regression and p is the number of samples. In a multiple regression model, the ordinary least squares approach (OLS)[35] is widely used for estimation[36] of parameters; in this work, we used the OLS method to estimate the parameters.

Ridge regression

The existence of near-linear relationships among the predictor variables is known as multicollinearity. Ridge regression is a modified version of linear regression for reducing multicollinearity among predictor variables in a multiple regression model.[37] Ridge regression utilises a modified loss function by adding a penalty equal to the square of the magnitude of the coefficients[27] as in Equation 4 to overcome this problem.

$$\frac{1}{p}\left\{\sum_{i=1}^{p}\left[Y_i - (\beta_0 + \sum_{j=1}^{n}\beta_j X_{ij})\right]^2\right\} + \lambda\sum_{j=0}^{n}\beta_j^2 \quad (4)$$

where the λ parameter regularises the penalty, ridge regression resembles linear regression; if $\lambda = 0$.

Lasso regression

Like ridge regression, lasso regression also evolved from linear regression to avoid multicollinearity problems. Instead of taking the coefficients' square as in ridge, lasso uses only magnitudes of the coefficients to penalise the cost function[28] as in Equation 5.

$$\frac{1}{p}\left\{\sum_{i=1}^{p}\left[Y_i - (\beta_0 + \sum_{j=1}^{n}\beta_j X_{ij})\right]^2\right\} + \lambda\sum_{j=0}^{n}\left|\beta_j\right| \quad (5)$$

k-Neighbours regression

The k-neighbours regression is developed based on the k-nearest neighbours' algorithm, a non-parametric method used for classification and regression.[29] The k-neighbours estimates the Y based on the local interpolation of the neighbourhood values determined while training. The size of the neighbourhood should be selected during the training phase using cross-validation. Since it is an instance-based learning method, k-neighbours adapts rapidly with new training data, enabling the algorithm to react quickly to variations in the training parameters during real-time operations.

Random forest regression

Random forest[30] incorporates several decision trees[38] and adds a layer of randomness to the bagging method. Random forests change how classification and regression trees are created and use a different bootstrap sample of the data for each tree.[39] The dependent variable's predicted value was obtained for the regression problem by averaging all decision trees' outputs. It builds as many more decision trees as possible, and each of these trees has high variance and low bias. However, averaging

all decision tree output to blend the results causes a final model with low bias and moderate variance, which reduces the overfitting problem and results in higher precision.

Support vector regression

Support vector regression (SVR)[31] is a machine learning technique that uses all of the significant features of the SVM and is highly robust against noise.[40] One of the critical benefits of SVR is that its computing complexity is independent of the input space's dimensionality. It also has a good generalisation potential and a high prediction accuracy.[41] The two main SVR types are ε-SVR[31] and v-SVR;[40] here, we used ε-SVR for strawberry firmness prediction because of its ability to control the error in the model.[42]

Partial least squares regression

PLSR integrates and generalises principal component analysis and multiple regression. It uses a series of independent variables or predictors to interpret a set of dependent variables. This prediction is made by generating a set of orthogonal variables called latent variables from the predictors with the most significant predictive power.[43] PLSR's usefulness comes from its ability to interpret noisy and collinear variables. The accuracy of the model parameters increases as the number of related variables and observations rises, which is a favourable property of PLSR.[23]

Processing pipeline

The acquired HSI data from the camera will go through a series of processing steps before reaching the regression models; they are preprocessing, normalised reflectance estimation, extraction of the mean spectrum from strawberry HSI and, finally, training and prediction using the selected regression method. The camera software performs the preprocessing of the data, including dark current removal, sensor corrections and radiometric calibration. During the normalisation process, the HSI data were transformed to normalised reflectance values between 0% and 100%, describing each pixel's spectral response, which was computed using the multilevel reference's known reflectance values, which were present in the scene during acquisition.

The spectral reflectance value can be determined as the reflected incident light ratio, as in Equation 6.

$$R(x,y,\lambda) = \frac{L_r(x,y,\lambda)}{L_i(x,y,\lambda)} \tag{6}$$

Here, the reflectance of wavelength λ at positions x, y is denoted as $R(x, y, \lambda)$, and the incident and reflected lights for wavelength λ at x, y are represented as L_i and L_r. The incident light can be determined by rearranging Equation 6, as shown below.

$$L_i(x,y,\lambda) = \frac{L_r(x,y,\lambda)}{R(x,y,\lambda)} \tag{7}$$

Equation 7 will be used to measure incident light across the field of view using the known reflectance of the reference target and reflected light intensities collected from the HSI sensor. We will use the incident light to estimate the entire HSI image's reflectance using Equation 6 since the device is a line scanner. After normalisation, the mean spectra of strawberries were calculated from a square-shaped region of interest (ROI) with a dimension of 200 × 200 pixels around the centre pixel of each strawberry. The extent is determined by the largest possible ROI that fits with all strawberries under investigation. As the final step, the obtained mean spectra were used for the regression.

A number of optimal wavelengths need to be selected from the HSI data to realise a multispectral imaging (MSI) system that can be used for potential real-time inspections. We used the successive projections algorithm (SPA)[32] to eliminate the redundant bands for dimensionality reduction. SPA is considered as a general, functional and robust recursive algorithm for solving near-separable non-negative matrix factorisation (NMF). The column of the input matrix $\mathbf{X}$ with the highest l_2 norm is selected at each step of the algorithm, and $\mathbf{X}$ is then updated by projecting each column into its orthogonal complement.[44] SPA has proved its capabilities in predicting fruit quality parameters;[45,46] hence, we used SPA to extract the critical bands from the strawberry spectra.

Training and evaluation

Each model will be calibrated and trained using 70% of the strawberries, and the remaining 30% will be used for prediction (testing); the training and test sets were selected randomly. The repeated K-fold technique with split count of ten and three repeats will train and cross-validate each regression model. Each regression model should go through parameter tuning while training, and the final model will be created based on the best parameters. Then the model will be used for the prediction and evaluation of the particular regression model. During the evaluation, five statistical parameters will be calculated to assess the efficiency of the established models. They are standard error of calibration

(*SEC*),[47] standard error of prediction (*SEP*),[47] mean absolute error (*MAE*),[48] mean squared error (*MSE*)[49] and coefficient of determination or R^2.[50] A decent model should have a high coefficient of determination (R^2) and a low *SEC*, *SEP*, *MSE* and *MAE*. Furthermore, the gap between *SEC* and *SEP* is a criterion for determining whether a model is suitable or not; lower is better. Besides calibration and prediction, the model metrics indicate overfitting; the model is potentially overfitting if it performs far better on the training set than on the test set.

Results and discussion

We implemented the entire processing pipeline using Python 3.6; the HSI data of 150 strawberries were acquired, processed and used for building regression models to predict the strawberry firmness. The mean spectrum obtained from all individual strawberry spectra was plotted and shown in Figure 3, along with the standard deviation (SD). The spectral profile contains 186 bands between 400 nm and 1000 nm with a 3.18 nm spectral resolution. The resulting spectral profile matches the previous studies; the sugar and water absorption bands are 840 nm and 960 nm, respectively,[51] in the near infrared (NIR) region. Anthocyanin and chlorophyll pigments, which reflect the fruit's colour characteristics, are found in regions around 535 nm and 680 nm.[1,52]

Analysis of variance (ANOVA) between the test and training firmness values gives a *p*-value of 0.47, which means that the training and test data sets seem statistically similar. Table 1 shows that k-neighbours and random forest regression accurately predicted the firmness values with very low and closer *SEP* and *SEC*, lower *MSE* and *MAE*, and higher R^2 values. Even though k-neighbours and random forest regression results were nearly similar, the k-neighbours performed marginally better. Hence, it is considered as the best regression according to the results we obtained. Linear regression has the lowest performance with *SEP* of 0.34, which is much closer to the previous research result.[25] This means that all the seven regression techniques performed pretty well in this experiment. The predicted vs measured firmness obtained for each regression and visualisation of the evaluation parameters are shown in Figure 4.

Each regression model was tuned for the best parameters using training data; for linear regression, the implementation was just plain OLS wrapped as a predictor for firmness. However, this model does not have any parameters for tuning; the final model was built using cross-validation of the training data. The prediction vs measured firmness for this model is displayed in Figure 4a. This model obtained an *MAE* of 0.14, *MSE* of 0.11, R^2 of 0.66 and *SEP* of 0.34 during evaluation, which is in an acceptable range and close to the calibration values.

In the ridge regression model, the tuning was done for the regularisation parameter λ, which determines the penalty. The grid search algorithm[53] was used for parameter optimisation and cross-validation on the training data set. The same algorithm was used for parameter tuning of lasso and SVR models. The optimal value obtained in this experiment was 0.02, which was selected from a grid of λ values between 0 and 1 at 0.01 step. The prediction vs

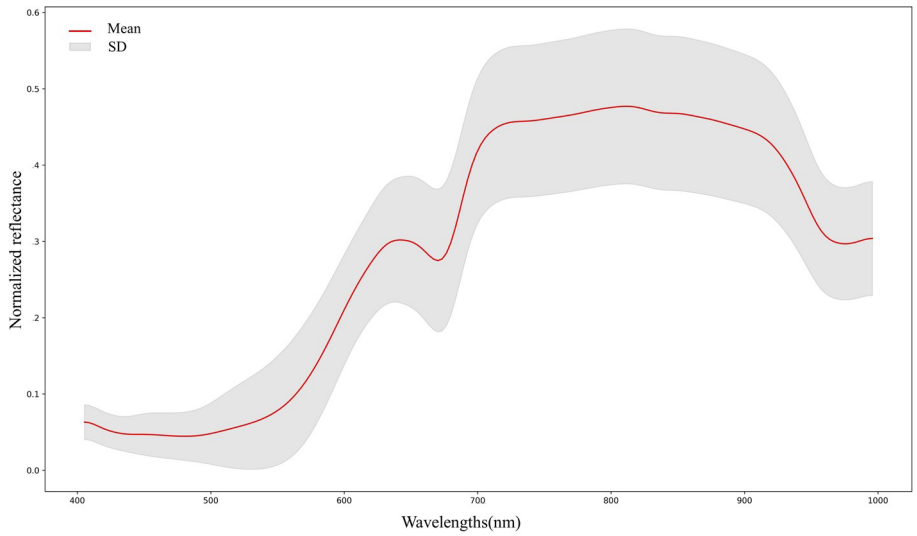


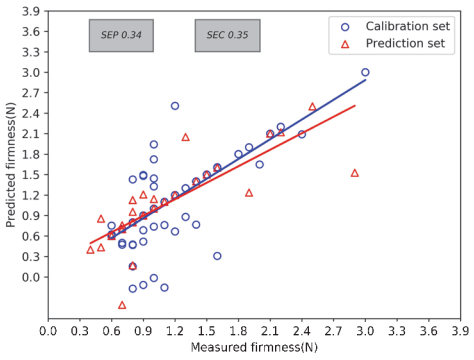
Figure 3. Mean spectrum of 150 strawberries with standard deviation.

Table 1. Results of all regression models with full spectral bands.

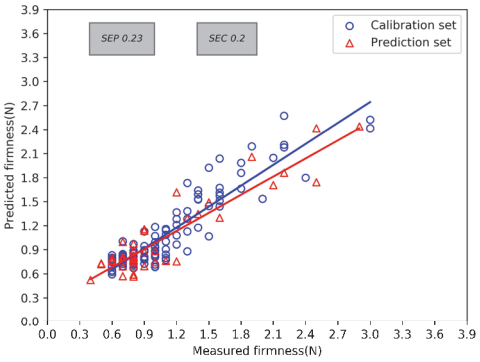
Parameters	Regression models													
	Linear		Ridge		Lasso		k-Neighbours		Random forest		SVR		PLSR	
	Cal	Pred	Cal	Pred	Cal	Pred	Cal	Pred	Cal	Pred	Cal	Pred	Cal	Pred
MAE	0.16	0.14	0.15	0.18	0.16	0.18	0.06	0.06	0.11	0.10	0.15	0.18	0.15	0.18
MSE	0.12	0.11	0.04	0.05	0.04	0.05	0.02	0.02	0.03	0.02	0.04	0.06	0.04	0.20
R^2	0.49	0.66	0.84	0.84	0.82	0.86	0.91	0.94	0.90	0.94	0.83	0.83	0.84	0.07
SEC	0.35	—	0.20	—	0.21	—	0.15	—	0.16	—	0.21	—	0.20	—
SEP	—	0.34	—	0.23	—	0.22	—	0.14	—	0.14	—	0.24	—	0.26

measured firmness for this model is displayed in Figure 4b. The ridge obtained a better *SEP* value of 0.23 with the help of regularisation. However, the performance was much better in comparison with linear regression having an *SEP* of 0.34. The lasso regression is also an extension of linear regression with L_1 norm as a regularisation parameter instead of L_2 in ridge regression. The optimisation procedure was the same as ridge and obtained 0.01 as the optimised parameter for λ. After optimisation, the lasso regression obtained an improved *SEP* of 0.22, a minor improvement from the ridge regression. The prediction vs measured firmness for the lasso model is displayed in Figure 4c. In ridge regression, the gap between the calibration and prediction values is higher than that of linear and lasso regression, indicating lower reliability. These differences can be quickly confirmed from Figure 4h.

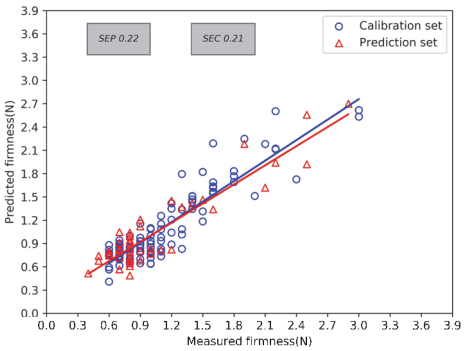
In k-neighbours regression, the prediction is made based on the closest local neighbours determined during the training phase. The number of neighbours "k" is the parameter that needs to be optimised to control the model's performance. The parameter optimisation was performed using random search for hyper-parameter optimisation;[54] unlike grid search, a fixed number of parameter settings are sampled from the specified distributions rather than all parameter values. The algorithm accepts a parameter that specifies how many different parameter settings should be attempted. After the parameter tuning using training data, we obtained the best *k* value as six, and the model obtained a *SEP* of 0.14 and is much closer to its *SEC* of 0.15. The other evaluation parameters obtained for this model during calibration and prediction are given in Table 1. The random forest regression also obtained nearly similar values to k-neighbours; however, this model has more parameters for tuning. The most significant parameters are the number of trees in the forest (estimators), a threshold number of attributes needed for splitting a node (max features), maximum levels in each decision tree (max depth), the minimum number of data points that can be added in a node before it is divided (min sample split), least quantity of data points permitted in a leaf node (min samples leaf) and whether bootstrap is enabled or not. Bootstrap decides the nature of sampling the data points; if enabled, sampling happens with replacement, otherwise without replacement. After the optimisation step, we obtained a model with an estimator count of 600, max features as the square root of the number of features, max depth of 30, min sample split of 2, min samples leaf of 4 and bootstrap as false. The optimisation was obtained using



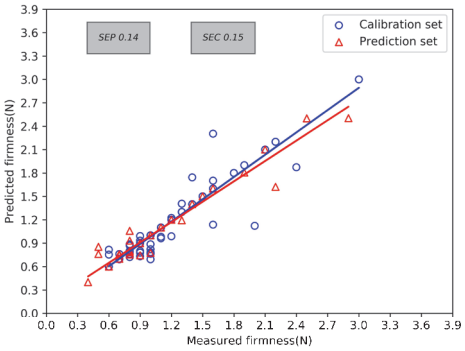
(a) Linear regression



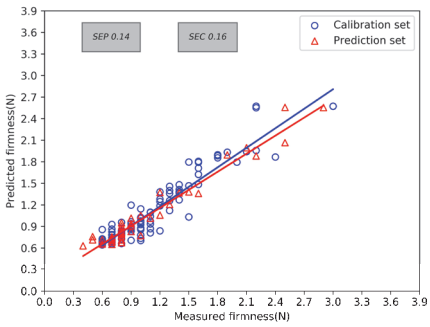
(b) Ridge regression



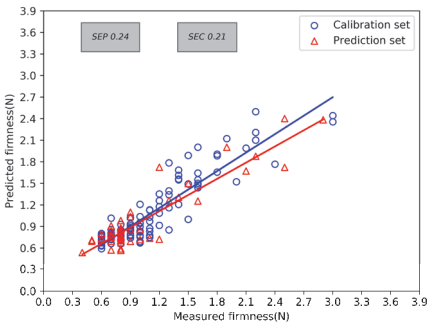
(c) Lasso regression



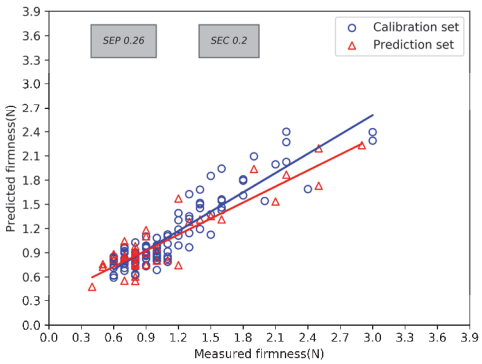
(d) k-Neighbours regression



(e) Random forest regression

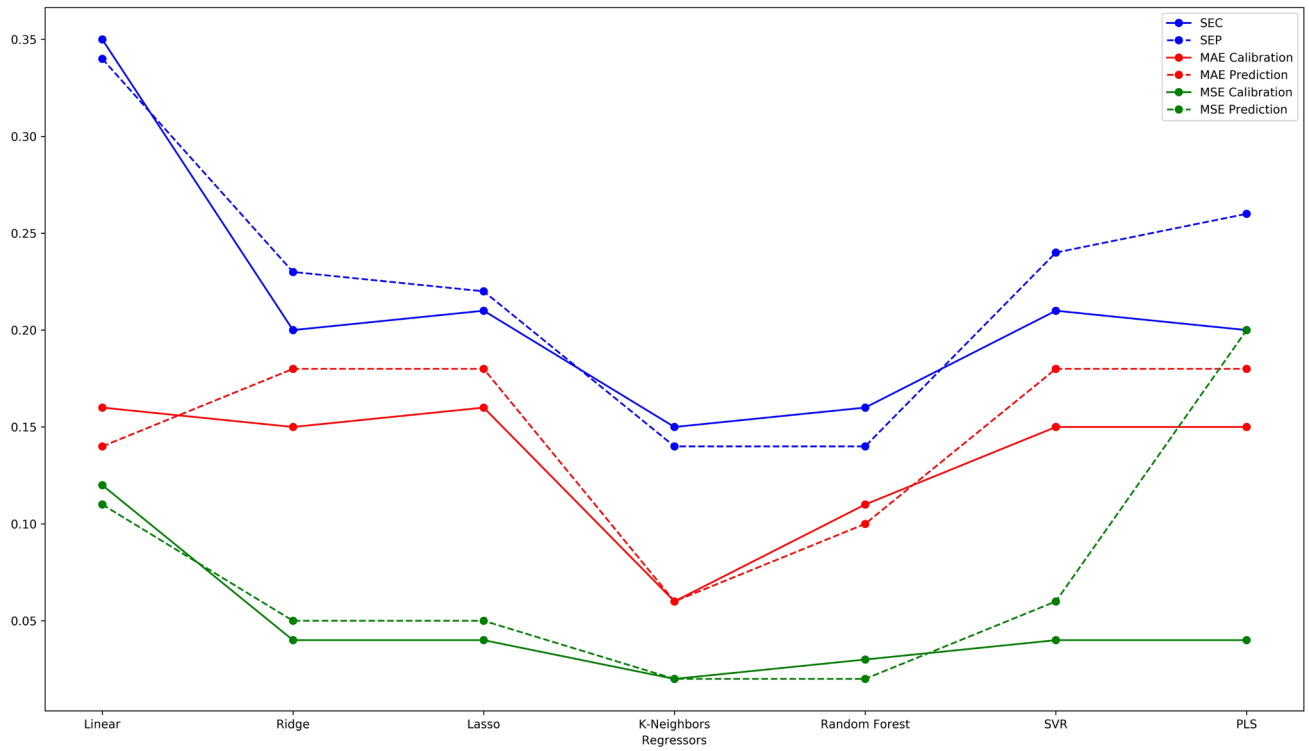


(f) SVR



(g) PLSR

Figure 4. Predicted vs measured firmness from all regression models used (a–h) and the evaluation parameter variation for each regression models (h).



(h) Evaluation parameter variation of regression models

Figure 4 (continued). Predicted vs measured firmness from all regression models used (a–h) and the evaluation parameter variation for each regression models (h).

the random search algorithm as in k-neighbours regression, and the implementation from the "sklearn" library[55] was used. This optimised model obtained an *SEP* of 0.14, the same as k-neighbours; however, this model has a slightly higher gap between *SEP* and *SEC* (gap of 0.02) than k-neighbours (0.01). The k-neighbours has zero gaps for *MSE* and *MAE* between calibration and prediction, which shows its reliability; however, the random forest model possesses a slight difference in *MSE* and *MAE* obtained during calibration and prediction. With these desirable parameters, the k-neighbours regression can be considered the best regression in this experiment; the prediction vs actual firmness plots for these two models are given in Figures 4d and 4e. Figure 5 shows the prediction map for strawberry firmness using the k-neighbours regression model having a measured firmness of 1.2 N, where the colourmap represents the predicted firmness variation.

The SVR model was successfully trained and evaluated, and yielded an *SEP* of 0.24 and an *SEC* of 0.21, and the prediction vs measured firmness plot is given in Figure 4f. The SVR model is tuned for parameters like kernel, the regularisation parameter *C*, kernel coefficient ϒ (gamma) and ε, which is the ε in the ε-SVR model. The kernel parameter was tested for three values—linear, RBF (radial basis function) and poly—and linear was identified as the most suitable kernel for this experiment. Also, we tested a range of values between 1.5 and 15 for *C*, 0 and 0.5 for ε and e^{-9} and e^{-4} for gamma using grid search optimisation. After optimisation, it was determined that 10,

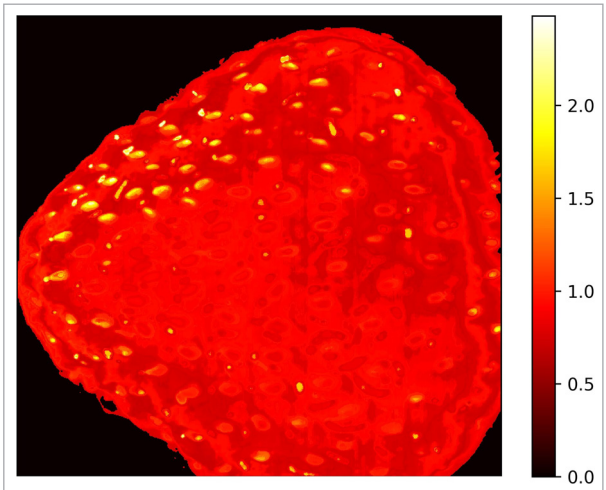


Figure 5. Firmness prediction map using the k-neighbours regression model.

0.1 and e^{-7} are the optimal values for *C*, ε and gamma, respectively.

The PLSR model was optimised against the parameter number of components to keep after dimensionality reduction. During optimisation, this parameter was varied between 1 and 10. It used cross-validation to observe and detect the best value and obtained six as the best number of components for this model. We used the NIPALS (Nonlinear Iterative Partial Least Squares)[56] algorithm to obtain singular vectors of the cross-covariance matrix and used the implementation from the "sklearn" library.[55] The results from PLSR are plotted in Figure 4g. This model obtained an *SEP* of 0.26 and an *SEC* of 0.20; the difference between *SEP* and *SEC* is significant, which indicates that this model is overfitted towards training data.

The k-neighbours model provided excellent *SEP* (0.14), *MSE* (0.02), *MAE* (0.06) and R^2 (0.94) values which are better than in the previous works reported in References 24–26. Mancini *et al.*[26] reported an R^2 of 0.54, which is significantly lower than that of k-neighbours. Sánchez *et al.*[24] reported the best *SEP* of 0.17, which is closer to our best model. Finally, Liu *et al.*[25] obtained an *SEP* of 0.375 and an R^2 value of 0.94; the *SEP* value is far from the best model obtained in this experiment with a similar R^2 value. However, because of its adaptability and ability to fit the data correctly without overfitting, k-neighbours is considered the best algorithm for firmness prediction. In the k-neighbours model, the data itself is a model that would be the reference for future prediction; it is very effective in improvising for random modeling on the available data; thus, it performed well in our case. However, the k-neighbours algorithm is susceptible to outliers because it uses the local neighbourhood values for prediction; it is also susceptible to imbalanced data. Hence while considering a more extensive data set for k-neighbours, it is essential to consider these weaknesses along with its strengths.

To develop an MSI system, a set of optimal wavelengths were chosen using the SPA algorithm to eliminate redundant details to realise HSI in prospective real-time analysis. The dataset contains 186 spectral bands ranging from 400 nm to 1000 nm, from which we extracted the 30 most important wavelengths using SPA. Out of these 30 bands, SPA identified band 128 (corresponding to wavelength 804.3 nm) as the most important one and 171 (corresponding to wavelength 941.6 nm) as the least important band. The k-neighbours model was re-established using these wavelengths and measured the evaluation parameters. The variation of these parameters against the various number of SPA bands is plotted in Table 2. According to the results, we can conclude that a multispectral system with three bands can provide a significant firmness prediction power. Those three bands are 804.3 nm, 552.2 nm and 667.1 nm, and the prediction vs measured value plot for this model is presented in Figure 6. The model with these three wavelengths produced the best evaluation matrices and has the lowest difference between calibration and prediction values. The 804 nm obtained from SPA as the most crucial wavelength, close to the 840 nm wavelength, is associated with the C–H bond related to carbohydrate content.[57] In addition to this, the following two SPAs were close to the wavelengths that determine the colour of the strawberries, 535 nm and 680 nm,[52] which can also be correlated with the firmness of the strawberry.

Conclusions

HSI data with various regression models were successfully calibrated, tuned and evaluated for strawberry firmness prediction. A total of 150 strawberries were used, and seven regression models were tested in this study. The k-neighbours model outperformed all other regression models with less error and overfitting. The relevant

Table 2. Evaluation of k-neighbours model on a various number of SPA bands.

Parameters	Number of SPA bands											
	1		3		5		10		20		30	
	Cal	Pred	Cal	Pred	Cal	Pred	Cal	Pred	Cal	Pred	Cal	Pred
MAE	0.19	0.47	0.15	0.20	0.14	0.22	0.15	0.21	0.15	0.22	0.15	0.24
MSE	0.32	0.34	0.07	0.07	0.04	0.09	0.04	0.07	0.04	0.09	0.05	0.09
R^2	0.22	0.08	0.82	0.79	0.84	0.74	0.82	0.78	0.83	0.74	0.79	0.70
SEC	0.43	—	0.21	—	0.19	—	0.20	—	0.20	—	0.22	—
SEP	—	0.59	—	0.27	—	0.29	—	0.27	—	0.29	—	0.32

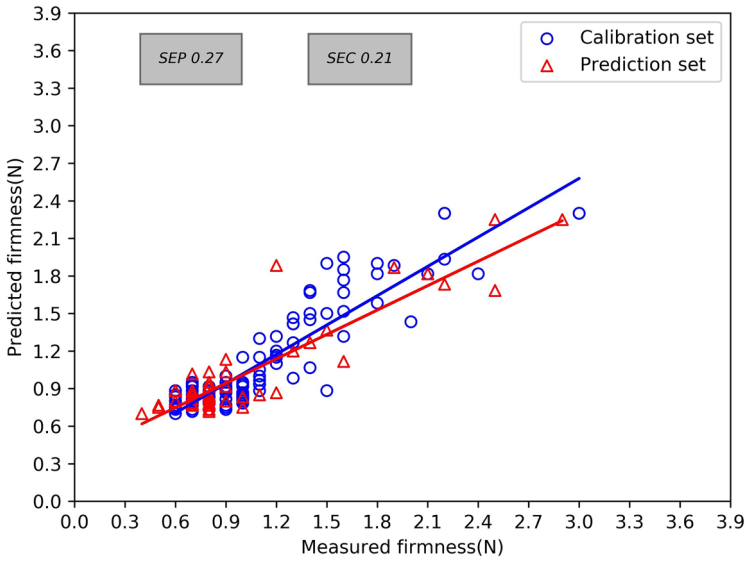


Figure 6. Predicted vs measured firmness of k-neighbours model with three SPA bands.

bands were identified from the original 186 bands using the SPA algorithm and re-established the k-neighbours model for varying SPA bands. We have identified that the k-neighbours model with the three most essential bands provided a comparable result of the entire band, which might be helpful in developing handheld, real-time applications. The potential future works will be to extend this research with more strawberries at different maturity levels and probing the potential of these models' prediction capability with other fruits.

Author contributions

B.M.D. and S.G. were responsible for the conceptualisation; B.M.D. worked on the methodology; software, B.M.D.; validation, B.M.D. and S.G.; data curation, B.M.D.; writing—original draft preparation, B.M.D.; writing—review and editing, S.G. and B.M.D.; supervision, S.G.; project administration, S.G.

Funding

This research received no external funding.

Conflicts of interest

The authors declare no conflict of interest.

Data availability

On fair request, the datasets produced during and/or analysed during the current study are available from the corresponding authors.

References

1. H. Wang, J. Peng, C. Xie, Y. Bao and Y. He, "Fruit quality evaluation using spectroscopy technology: a review", *Sensors* **15(5),** 11889 (2015). https://doi.org/10.3390/s150511889
2. M.G.H. Stec, J.A. Hodgson, E.A. Macrae and C.M. Triggs, "Role of fruit firmness in the sensory evaluation of kiwifruit (*Actinidia deliciosa* cv Hayward)", *J. Sci. Food Agric.* **47(4),** 417 (1989). https://doi.org/10.1002/jsfa.2740470404
3. M. Nagata, J.G. Tallada, T. Kobayashi, Y. Cui and Y. Gejima, "Predicting maturity quality parameters of strawberries using hyperspectral imaging", *ASAE Annu. Int. Meet. 2004* 043033 (2004). https://doi.org/10.13031/2013.16704
4. J. Lado, E. Vicente, A. Manzzioni and G. Aresb, "Application of a check-all-that-apply question for the evaluation of strawberry cultivars from a breeding program", *J. Sci. Food Agric.* **90(13),** 2268 (2010). https://doi.org/10.1002/jsfa.4081
5. A. Døving and F. Måge, "Methods of testing strawberry fruit firmness", *Acta Agric. Scand.*

Sect. B Soil Plant Sci. **52(1),** 43 (2002). https://doi.org/10.1080/090647102320260035

6. F. Duprat, M. Grotte, E. Pietri and D. Loonis, "The acoustic impulse response method for measuring the overall firmness of fruit", *J. Agric. Eng. Res.* **66(4),** 251 (1997). https://doi.org/10.1006/jaer.1996.0143
7. S.K. Jha, S. Sethi, M. Srivastav, A.K. Dubey, R.R. Sharma, D.V.K. Samuel and A.K. Singh, "Firmness characteristics of mango hybrids under ambient storage", *J. Food Eng.* **97(2),** 208 (2010). https://doi.org/10.1016/j.jfoodeng.2009.10.011
8. J. Qin, K. Chao, M.S. Kim, R. Lu and T.F. Burks, "Hyperspectral and multispectral imaging for evaluating food safety and quality", *J. Food Eng.* **118(2),** 157 (2013). https://doi.org/10.1016/j.jfoodeng.2013.04.001
9. J. Steinbrener, K. Posch and R. Leitner, "Hyperspectral fruit and vegetable classification using convolutional neural networks", *Comput. Electron. Agric.* **162,** 364 (2019). https://doi.org/10.1016/j.compag.2019.04.019
10. B.M. Devassy and S. George, "Contactless classification of strawberry using hyperspectral imaging", in *CEUR Workshop Proceedings*, **2688** (2020). https://dblp.org/rec/conf/cvcs/DevassyG20
11. M.A. Calin, S.V. Parasca, D. Savastru and D. Manea, "Hyperspectral imaging in the medical field: present and future", *Appl. Spectrosc. Rev.* **49(6),** 435 (2014). https://doi.org/10.1080/05704928.2013.838678
12. G. Lu and B. Fei, "Medical hyperspectral imaging: a review", *J. Biomed. Opt.* **19(1),** 010901 (2014). https://doi.org/10.1117/1.jbo.19.1.010901
13. B.M. Devassy and S. George, "Forensic analysis of beverage stains using hyperspectral imaging", *Sci. Rep.* **11(1),** 1 (2021). https://doi.org/10.1038/s41598-021-85737-x
14. B.M. Devassy, S. George and J.Y. Hardeberg, "Comparison of ink classification capabilities of classic hyperspectral similarity features", in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, p. 25 (2019).
15. H. Deborah, S. George and J.Y. Hardeberg, "Pigment mapping of *The Scream* (1893) based on hyperspectral imaging", *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **8509 LNCS,** 247 (2014). https://doi.org/10.1007/978-3-319-07998-1_28
16. S. George and J.Y. Hardeberg, "Ink classification and visualisation of historical manuscripts: application of hyperspectral imaging", in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, **2015-Novem**, p. 1131 (2015).
17. B. Melit Devassy, S. George and P. Nussbaum, "Unsupervised clustering of hyperspectral paper data using t-SNE", *J. Imaging* **6(5),** 29 (2020). https://doi.org/10.3390/jimaging6050029
18. P. Rajkumar, N. Wang, G. Elmasry, G.S.V. Raghavan and Y. Gariepy, "Studies on banana fruit quality and maturity stages using hyperspectral imaging", *J. Food Eng.* **108(1),** 194 (2012). https://doi.org/10.1016/j.jfoodeng.2011.05.002
19. C. Yang, W.S. Lee and P. Gader, "Hyperspectral band selection for detecting different blueberry fruit maturity stages", *Comput. Electron. Agric.* **109,** 23 (2014). https://doi.org/10.1016/j.compag.2014.08.009
20. R. Lu and Y. Peng, "Hyperspectral scattering for assessing peach fruit firmness", *Biosyst. Eng.* **93(2),** 161 (2006). https://doi.org/10.1016/j.biosystemseng.2005.11.004
21. M. Tranmer, J. Murphy, M. Elliot and M. Pampaka, "Multiple linear regression (2nd edition)", *Cathie Marsh Inst. Work. Pap.* **(01)**, 59 (2020).
22. J.G. Tallada, M. Nagata and T. Kobayashi, "Non-destructive estimation of firmness of strawberries (*Fragaria × ananassa duch.*) using NIR hyperspectral imaging", *Environ. Control Biol.* **44(4),** 245 (2006). https://doi.org/10.2525/ecb.44.245
23. S. Wold, M. Sjöström and L. Eriksson, "PLS-regression: a basic tool of chemometrics", *Chemometr. Intell. Lab. Syst.* **58(2),** 109 (2001). https://doi.org/10.1016/S0169-7439(01)00155-1
24. M.T. Sánchez, M.J. De La Haba, M. Benítez-López, J. Fernández-Novales, A. Garrido-Varo and D. Pérez-Marín, "Non-destructive characterization and quality control of intact strawberries based on NIR spectral data", *J. Food Eng.* **110(1),** 102 (2012). https://doi.org/10.1016/j.jfoodeng.2011.12.003
25. C. Liu, W. Liu, X. Lu, F. Ma, W. Chen, J. Yang and L. Zheng, "Application of multispectral imaging to determine quality attributes and ripeness stage in strawberry fruit", *PLoS One* **9(2),** (2014). https://doi.org/10.1371/journal.pone.0087818
26. M. Mancini, L. Mazzoni, F. Gagliardi, F. Balducci, D. Duca, G. Toscano, B. Mezzetti and F. Capocasa, "Application of the non-destructive NIR technique for the evaluation of strawberry fruits quality parameters", *Foods* **9(4),** 441 (2020). https://doi.org/10.3390/foods9040441

27. A.E. Hoerl and R.W. Kennard, "Ridge regression: applications to nonorthogonal problems", *Technometrics* **12(1),** 69 (1970). https://doi.org/10.1080/00401706.1970.10488635

28. R. Tibshirani, "Regression shrinkage and selection via the lasso", *J.R. Stat. Soc. Ser. B* **58(1),** 267 (1996). https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

29. N.S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression", *Am. Stat.* **46(3),** 175 (1992). https://doi.org/10.1080/00031305.1992.10475879

30. L. Breiman, "Random forests", *Mach. Learn.* **45(1),** 5 (2001). https://doi.org/10.1023/A:1010933404324

31. H. Drucker, C.J.C. Surges, L. Kaufman, A. Smola and V. Vapnik, "Support vector regression machines", in *Advances in Neural Information Processing Systems*, p. 155 (1997).

32. M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame and V. Visani, "The successive projections algorithm for variable selection in spectroscopic multicomponent analysis", *Chemometr. Intell. Lab. Syst.* **57(2),** 65 (2001). https://doi.org/10.1016/S0169-7439(01)00119-8

33. S. Weisberg, *Applied Linear Regression*, 3rd Edn. Wiley (2005). https://doi.org/10.1002/0471704091

34. S. Chattefuee and A.S. Hadi, *Regression Analysis by Example*, 4th Edn. Wiley (2006). https://doi.org/10.1002/0470055464

35. B.G.D. Hutcheson, *Ordinary Least-Squares Regression* (2011).

36. G.K.F. Tso and K.K.W. Yau, "Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks", *Energy* **32(9),** 1761 (2007). https://doi.org/10.1016/j.energy.2006.11.010

37. L. Vinet and A. Zhedanov, "A 'missing' family of classical orthogonal polynomials", *J. Phys. A: Math. Theor.* **44,** 085201 (2011). https://doi.org/10.1088/1751-8113/44/8/085201

38. A. Priyam, R. Gupta, A. Rathee and S. Srivastava, "Comparative analysis of decision tree classification algorithms", *Int. J. Curr. Eng. Technol.* **3(2),** 334 (2013). http://inpressco.com/comparative-analysis-of-decision-tree-classification-algorithms/

39. A. Liaw and M. Wiener, "Classification and regression by randomForest", *R News* **2(3),** 18 (2002).

40. B. Schölkopf, A.J. Smola, R.C. Williamson and P.L. Bartlett, "New support vector algorithms", *Neural Comput.* **12(5),** 1207 (2000). https://doi.org/10.1162/089976600300015565

41. M. Awad and R. Khanna, "Support vector regression", *Efficient Learning Machines*. Apress, pp. 67–80 (2015). https://doi.org/10.1007/978-1-4302-5990-9_4

42. M. Tan, X. Song, X. Yang and Q. Wu, "Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: acomparative study", *J. Nat. Gas Sci. Eng.* **26,** 792 (2015). https://doi.org/10.1016/j.jngse.2015.07.008

43. H. Abdi, "Partial least squares regression", in *Encyclopedia of Measurement and Statistics*, Ed by N.J. Salkind. Sage (2007).

44. N. Gillis, "Successive nonnegative projection algorithm for robust nonnegative blind source separation", *SIAM J. Imaging Sci.* **7(2),** 1420 (2014). https://doi.org/10.1137/130946782

45. F. Liu and Y. He, "Application of successive projections algorithm for variable selection to determine organic acids of plum vinegar", *Food Chem.* **115(4),** 1430 (2009). https://doi.org/10.1016/j.foodchem.2009.01.073

46. Y. Sun, X. Gu, K. Sun, H. Hu, M. Xu, Z. Wang, K. Tu and L. Pan, "Hyperspectral reflectance imaging combined with chemometrics and successive projections algorithm for chilling injury classification in peaches", *LWT - Food Sci. Technol.* **75,** 557 (2017). https://doi.org/10.1016/j.lwt.2016.10.006

47. M. Forina, G. Drava, C. Armanino, R. Boggia, S. Lanteri, R. Leardi, P. Corti, P. Conti, R. Giangiacomo, C. Galliena, R. Bigoni, I. Quartari, C. Serra, D. Ferri, O. Leoni and L. Lazzeri, "Transfer of calibration function in near-infrared spectroscopy", *Chemometr. Intell. Lab. Syst.* **27(2),** 189 (1995). https://doi.org/10.1016/0169-7439(95)80023-3

48. C.J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance", *Clim. Res.* **30(1),** 79 (2005). https://doi.org/10.3354/cr030079

49. Z. Wang and A.C. Bovik, "Mean squared error: lot it or leave it? A new look at signal fidelity measures", *IEEE Signal Process. Mag.* **26(1),** 98 (2009). https://doi.org/10.1109/MSP.2008.930649

50. O. Renaud and M.P. Victoria-Feser, "A robust coefficient of determination for regression", *J. Stat. Plan. Inference* **140(7),** 1852 (2010). https://doi.org/10.1016/j.jspi.2010.01.008

51. Y. Sun, Y. Liu, H. Yu, A. Xie, X. Li, Y. Yin and X. Duan, "Non-destructive prediction of moisture content and freezable water content of purple-fleshed sweet potato slices during drying process using hyperspectral imaging technique", *Food Anal. Meth.* **10(5),** 1535 (2017). https://doi.org/10.1007/s12161-016-0722-0

52. X.Q. Yue, Z.Y. Shang, J.Y. Yang, L. Huang and Y.Q. Wang, "A smart data-driven rapid method to recognize the strawberry maturity", *Inf. Process. Agric.* **7(4),** 575 (2020). https://doi.org/10.1016/j.inpa.2019.10.005

53. Q. Huang, J. Mao and Y. Liu, "An improved grid search algorithm of SVR parameters optimization", in *International Conference on Communication Technology Proceedings*, *ICCT*, p. 1022 (2012).

54. J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization", *J. Mach. Learn. Res.* **13(1),** 281–305 (2012). https://jmlr.org/papers/v13/bergstra12a.html

55. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, "Scikit-learn: machine learning in python", *J. Mach. Learn. Res.* **12(85),** 2825–2830 (2011). https://jmlr.org/papers/v12/pedregosa11a.html

56. H. Wold, "11-path models with latent variables: the NIPALS approach", in *International Perspectives on Mathematical and Statistical Modeling*, p. 307 (1975).

57. M. Kamruzzaman, Y. Makino and S. Oshita, "Parsimonious model development for real-time monitoring of moisture in red meat using hyperspectral imaging", *Food Chem.* **196,** 1084–1091 (2016). https://doi.org/10.1016/j.foodchem.2015.10.051