applied sciences

MDPI

Article

Exploration of Machine Learning Algorithms for pH and Moisture Estimation in Apples Using VIS-NIR Imaging

Erhan Kavuncuoğlu [1], Necati Çetin [2,*], Bekir Yildirim [3], Mohammad Nadimi [4] and Jitendra Paliwal [4,*]

[1] Department of Computer Technologies, Gemerek Vocation School, Cumhuriyet University, Sivas 58140, Turkey; erhankav@gmail.com
[2] Department of Agricultural Machinery and Technologies Engineering, Faculty of Agriculture, Ankara University, Ankara 06110, Turkey
[3] Department of Textile Engineering, Faculty of Engineering, Erciyes University, Kayseri 74110, Turkey
[4] Department of Biosystems Engineering, University of Manitoba, Winnipeg, MB R3T 5V6, Canada; mohammad.nadimi@umanitoba.ca
* Correspondence: necati.cetin@ankara.edu.tr (N.Ç.); j.paliwal@umanitoba.ca (J.P.)

Abstract: Non-destructive assessment of fruits for grading and quality determination is essential to automate pre- and post-harvest handling. Near-infrared (NIR) hyperspectral imaging (HSI) has already established itself as a powerful tool for characterizing the quality parameters of various fruits, including apples. The adoption of HSI is expected to grow exponentially if inexpensive tools are made available to growers and traders at the grassroots levels. To this end, the present study aims to explore the feasibility of using a low-cost visible-near-infrared (VIS-NIR) HSI in the 386–1028 nm wavelength range to predict the moisture content (MC) and pH of Pink Lady apples harvested at three different maturity stages. Five different machine learning algorithms, viz. partial least squares regression (PLSR), multiple linear regression (MLR), k-nearest neighbor (kNN), decision tree (DT), and artificial neural network (ANN) were utilized to analyze HSI data cubes. In the case of ANN, PLSR, and MLR models, data analysis modeling was performed using 11 optimum features identified using a Bootstrap Random Forest feature selection approach. Among the tested algorithms, ANN provided the best performance with R (correlation), and root mean squared error ($RMSE$) values of 0.868 and 0.756 for MC and 0.383 and 0.044 for pH prediction, respectively. The obtained results indicate that while the VIS-NIR HSI promises success in non-destructively measuring the MC of apples, its performance for pH prediction of the studied apple variety is poor. The present work contributes to the ongoing research in determining the full potential of VIS-NIR HSI technology in apple grading, maturity assessment, and shelf-life estimation.

Keywords: hyperspectral imaging; apple; pH; moisture content; machine learning

Citation: Kavuncuoğlu, E.; Çetin, N.; Yildirim, B.; Nadimi, M.; Paliwal, J. Exploration of Machine Learning Algorithms for pH and Moisture Estimation in Apples Using VIS-NIR Imaging. *Appl. Sci.* **2023**, *13*, 8391. https://doi.org/10.3390/app13148391

Academic Editors: Salik Khanal and Vitor Filipe

Received: 29 May 2023
Revised: 7 July 2023
Accepted: 17 July 2023
Published: 20 July 2023

1. Introduction

Apple is one of the most widely grown fruits around the world. Over the past decade, there has been a remarkable increase in apple production worldwide. Between 2010 and 2020, worldwide apple production grew from 71.19 million metric tons to 86.44 million metric tons [1]. This edible fruit is a good source of antioxidants such as ascorbic acid and polyphenols, which are known to improve human health. Apple is also high in vitamins, minerals, and sugars [2–4]. The fruit grade/quality is determined by various factors, including mass, pH, firmness, soluble solids content (SSC), color, moisture, and internal browning [5]. Most of these parameters are also critical in determining apple ripeness and/or optimal harvest times. Among them, moisture content (MC) is the most important quality attribute that directly affects the fruit's shelf life [6].

Unfortunately, the conventional methods for evaluating the aforementioned quality traits are destructive, subjective, time-consuming, and/or prone to operational errors. Therefore, identifying intelligent and non-destructive alternative technologies for assessing apple quality is of great interest to the industry [7]. Over the past decade, hyperspectral imaging (HSI) has shown promising performance as a reliable tool for the safety and quality assessment of agricultural products [8,9]. This non-destructive method combines the features of traditional imaging and spectroscopy to simultaneously analyze spatial and spectral information of a sample, making it an invaluable tool for assessing quality indicators of fruits, vegetables, legumes, oilseeds, and grains [7]. For example, scholars previously explored the capability of HSI in assessing the MC of various fruits such as strawberries [10], mango [11], tomatoes [12], potatoes [13], mushrooms [14], and persimmons [15]. Moreover, scientists implemented HSI for pH prediction of various fruits such as grapes [16], apples [17], peaches [18], kiwifruit [19], cherries [20] and strawberries [21].

Despite several research studies on the use of HSI to monitor fruit quality, one of the main challenges in implementing this technology in the industry has been the enormity of the datasets and the cumbersome analysis that is required to glean useful information from them. One potential solution to overcome this challenge may be the implementation of state-of-the-art machine learning algorithms [22–24]. Over the past few years, machine learning algorithms have become increasingly popular in classification and prediction research, with techniques such as k-nearest neighbor (kNN), artificial neural networks (ANN), decision trees (DT), and convolutional neural networks (CNN) being widely employed [25–32].

The present work aims to assess apples' pH and MC variations at different maturity stages. Unlike the majority of hyperspectral data analysis models that only utilize partial least squares regression (PLSR) and/or multiple linear regression (MLR), herein, we explore the performance of three additional machine learning algorithms (i.e., kNN, DT, and ANN) in predicting the industry-accepted quality parameters to contribute to the ongoing effort in resolving data analysis challenge of hyperspectral data. Considering we previously demonstrated the capability of visible-near-infrared (VIS-NIR) HSI technology for apple firmness, SSC [33], ripening levels [34], phenolic content, antioxidant activity, and ascorbic acid prediction [4,35], the present work contributes to revealing the full potential of VIS-NIR HSI in apple grading, maturity assessment, and shelf life estimation. Such a tool has the potential to enhance sustainable apple production through improved quality assessment, waste reduction, and increased energy efficiency in data processing.

2. Material and Methods

2.1. Apple Samples

One hundred Pink Lady apples harvested at three different stages of maturity were selected (in total, 300 apples) for HSI acquisition. The first harvest date was 10th October 2019 (maturity stage 1), and the subsequent harvesting took place on the 7th day (maturity stage 2) and 14th day (maturity stage 3) after that. The apples selected for spectral data collection were sound without any damage or deterioration on their exterior surface. Their soundness was confirmed via visual inspection by a panel of experts. The collected apples were individually labeled and stored at 4 °C prior to being imaged. Hyperspectral image acquisition of all samples was completed within two weeks of harvesting.

2.2. Standard Measurements for Quality Attributes

2.2.1. pH Measurement

In order to determine apple pH values, sample juices were extracted and analyzed using a pH meter (Hanna HI 2002-02, Woonsocket, RI, USA).

2.2.2. Moisture Content Measurement

For MC measurement, 100 g of flesh fruit from 100 different samples for each maturity stage was placed in an oven (Memmert UN55, Schwabach, Germany) at 70 °C for 48 h. MCs (wet basis) were obtained by using Equation (1) [36].

$$M_c = \frac{W_i - W_f}{W_i} \times 100 \quad (1)$$

where M_c represents moisture content (%), W_i indicates the initial mass (g), and W_f represents the product's final mass (g).

2.3. Data Collection

2.3.1. Hyperspectral Imaging System

The apple samples were imaged using a push-broom hyperspectral camera (PIKA-L, Resonon Inc., Bozeman, MT, USA). The HSI system was comprised of five components, including a PIKA HSI camera, a supporting tower, a motion platform, an illumination source, and a system controller. This device collected reflectance data across 300 spectral bands, ranging from 386–1028 nm wavelengths, producing a spectral resolution of 2.1 nm, with a digital yield of 12 bits. The target lens, with a focal length of 17 mm, was specifically tailored for NIR and VIS-NIR spectra. Lighting conditions were maintained consistent with the help of four 15 W 12 V bulbs positioned symmetrically around the lens. The camera was set up 50 cm over the linear movement stage, generating a spatial resolution of approximately 50 pixels per square millimeter.

Prior to initiating data collection, a dark calibration process was carried out with several dark frames captured while the lens cover was kept shut. These readings served to counteract the dark current noise during actual measurements. A Teflon piece (K-Mac Plastics, Wyoming, MI, USA) was utilized to ensure white balance. The illumination system was turned on for a half-hour prior to capturing the images to bring the system to a stable state. The spectral data were then derived from the processed images [33]

2.3.2. Image Analysis

Image segmentation and selection of a region of interest (ROI) is an important step in the analysis of HSI data. Automatic thresholding (Otsu's method) was initially performed to segment the apples from their background, but the performance was unreliable. Therefore, an alternative approach based on the calculation of standard deviations of the spectral reflectance values of all pixels was adopted. The details of this segmentation procedure can be found in [33].

2.4. Data Pre-Processing, Feature Selection and Cross-Validation

All measured pH and MC values were normalized for training and testing. Spectral reflectance, pH, and MC data were normalized using the Z-score normalization method according to Equation (2):

$$x_i' = \frac{x_i - \overline{x}}{\sigma_x} \tag{2}$$

where x_i, $\overline{x}$, and σ_x represent the value of the ith observation, mean, and the standard deviation of the x-variable, respectively.

The feature selection process in this study was conducted using embedded methods, as described in [33]. Specifically, the Bootstrap Random Forest technique was employed as a machine learning method for bagging ensemble learning. This involved training multiple Random Forest trees on different subsets of observations using bootstrap sampling, where each tree is trained on a different subset of observations. The remaining observations, known as the out of bag (OOB) sample, are then used to estimate the model's performance. The final model predictions are generated by averaging the outcomes of all the trees. This process is designed to reduce variance and improve the model's accuracy.

To identify the most important features for predicting pH and MC values, 100 different spectrum features were extracted for each wavelength at each maturity stage. The features that best estimated each output value were then selected. Model performance was evaluated using k-fold cross-validation, where k = 10 owing to the 10 subsets in our dataset. The training set was split into 10 sub-sets, with one subset reserved for testing and the other nine used for training in each iteration. This process was repeated 10 times to ensure the robustness of the results [37].

2.5. Machine Learning Algorithms and Statistical Methods

This section describes the models used for analyzing the VIS-NIR HSI data. Python 3.11 software was employed to make predictions using machine learning algorithms on a computer equipped with a core i5 CPU running at 3.2 GHz and 12 GB of memory. The dataset consisted of measurements for 100 different apples across the three maturity stages. The entire dataset was randomly divided into 70% spectra for training and 30% for testing purposes for all models.

2.5.1. Artificial Neural Network

ANN is a machine-learning tool that can be utilized for pattern recognition. Herein, multilayer feed-forward neural network (MFFNN) structure was used as an ANN tool to predict the desired parameters. A typical MFFNN structure involves an input layer, multiple hidden layers, and an output layer. Our model architectures consisted of one-to-three hidden layers. The number of neurons ranged from 5 to 90 in each hidden layer to determine the most reliable ANN structure. Different activation functions, viz. hyperbolic tangent (*tanh*), logarithmic sigmoid (*Logsig*), linear (*Purelin*), and Gaussian, were used to update the network's weights.

The developed model included an input layer fed by 11 features, two hidden layers, and an output layer. The first hidden layer had 10 computational units (neurons) with the sigmoid activation function. The second hidden layer had 70 neurons with the linear activation function. The ANN model had a layer configuration of 11-10-70-2. However, only the best-performing model's outcomes have been presented in the results section.

A total of 500 epochs were used for the analysis, and the activation functions *tanh-tanh* were applied to both pH and MC output parameters. Analyses were conducted using a total of 500 epochs. The *tanh-tanh* activation function was applied to both output parameters to capture the complex non-linear patterns within the data: MC and pH. This activation function maps input values within a wide range to a bounded output between −1 and +1, offering non-linearity, continuous differentiability, and confined outputs.

The learning rate, a crucial factor in ANN models, determines the pace at which the model assimilates information and adjusts its parameters during training. It remarkably impacts the model's convergence and ability to find an optimal solution. A trial-and-error approach was used to identify the ideal learning rate for the neural network. Different learning rates (0.001, 0.01, 0.1, 0.2, and 0.3) were explored to pinpoint the rate that led to the most reliable performance. After training iterations, it was found that a learning rate of 0.01 provided the best performance. This value expedited smooth convergence during the training process and resulted in superior prediction accuracy.

2.5.2. Decision Tree

As a supervised learning algorithm, a DT works on a set of tree-like decision rules and their various consequences to classify/predict the desired value of a dependent variable [38]. In fact, the algorithm employs conditional control rules to make a prediction. The DT regression mechanism is described in detail elsewhere [39]. The numbers of split trees in the DT models were 2 and 5 for pH and MC, respectively.

2.5.3. k-Nearest Neighbor Algorithm

The kNN model is an important non-parametric supervised proximity-based machine learning predictive model. Initially, the algorithm considers all of the observation values as a cluster. The clusters then progressively combine to create new clusters. In kNN regression, the dependent variable is usually obtained by averaging the dependent values of k nearest neighbors in the training models. The Euclidean distance method is commonly implemented to identify the closest neighbors [40,41]. More details on the principles of kNN regression can be found elsewhere [42]. In this work, k was set to 10 for both pH and MC predictions.

2.5.4. Multiple Linear Regression

MLR is a common regression method to predict a linear relationship between dependent and independent variables. This technique implements several independent variables to predict the outcome of a dependent variable by minimizing the difference between the predictions and the actual values of the target variable. In this work, Equation (3) [43] was used to estimate a dependent (response) variable (y) using the selected features (see Section 2.4) as independent variables.

$$y = a + \sum_{i=1}^{n} b_i x_i + \varepsilon \tag{3}$$

where ε represents the error, a is an intercept, b_i is a regression coefficient, and x_i is a predictor variable. More details on the principles of MLR regression can be found elsewhere [33].

2.5.5. Partial Least Squares Regression

PLSR is a well-known multivariate regression technique for calibrating the NIR data [44]. The method takes into account the structure of both dependent and independent variables. While similar to MLR, the PLSR model can predict linear relations between dependent and independent variables; it identifies regression coefficients differently than MLR. In PLSR, the dependent and independent variables will project into latent structures in an iterative process. The latent structure with the highest variability for the dependent variable is extracted and explained by a latent structure of the independent variable to identify the best predictive model. More details on the principles of PLSR regression are provided elsewhere [44].

2.6. Performance Assessment of the Models and Statistical Analyses

The performance of models was evaluated with the use of correlation coefficient (R), root mean square error ($RMSE$), mean absolute error (MAE), and mean absolute percentage error ($MAPE$), as shown below:

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(M_i - \dot{M})\,(E_i - \dot{E})}{S_M\, S_E} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (E_i - M_i)^2}{n}} \tag{5}$$

$$MAE = \sum_{i=1}^{n} \frac{|\ E_i - M_i\ |}{n} \tag{6}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{E_i - M_i}{E_i} \right| \times 100 \tag{7}$$

where n is the number of data instances, M_i measured values, E_i predicted values, M mean measured values, $\dot{E}$ mean predicted value, S_M measured target values sum, and S_E predicted target values sum. The R was analyzed to assess the reliability of model prediction [45].

3. Results

The pH and MC of the apples were measured to be 3.40 ± 0.05 (mean ± standard deviation) and 87.82 ± 1.43% (w.b.), respectively.

Figure 1 shows the mean NIR spectra of apple samples at different maturity stages. It is apparent that the reflectance values are higher in the red region (630 nm) for the maturity stage 3. One can note a transition occurring at about 610 nm. Such observation is reasonable as the color of the apples turns redder as the samples mature. In the meantime, the apples from the maturity stage 1 (less matured samples) have a stronger peak around the green region (550 nm).

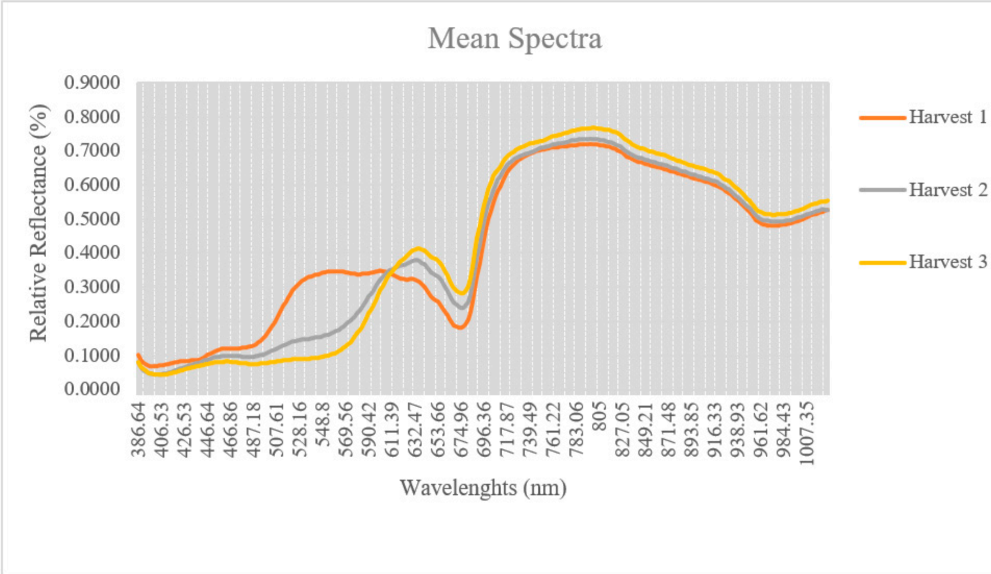


Figure 1. Mean spectra of different maturity stages.

The feature selection algorithms ranked features according to their contribution to the prediction of the target variables. Figure 2a shows the 11 best corresponding features for MC and pH. Among them, a total of 11 features (10 wavelengths + 1 maturity stage) were identified as optimum features for the simultaneous detection of MC and pH for all of our subsequent analyses. These features include 466.86, 468.89, 472.94, 666.43, 689.93, 692.07, 694.22, 696.36, 702.80, 1028.06 nm, and "Maturity stages".

The 11 features identified were used as inputs in the ANN, with an input layer fed by 11 features, two hidden layers, and one output layer. The first and second hidden layers had 10 and 70 neurons, resulting in an overall ANN structure of 11-10-70-2 (Figure 2b).

Table 1 shows the performance of the various predictive model for the testing set. The performance of the MC predictors was promising, with *R*-values in the range of 0.844–0.868, with the best performance achieved under ANN. However, the developed model proved to be unreliable in predicting the pH value, with *R*-values ranging from 0.190 to 0.383. Despite this, the ANN model still delivered the best performance.

Table 1. Performance results of the machine learning algorithms.

Model	Outputs	*R*	*RMSE*	*MAE*	*MAPE*
ANN *	pH	0.383	0.044	0.035	0.010
	MC	0.868	0.756	0.578	0.007
DT **	pH	0.307	0.046	0.036	0.011
	MC	0.844	0.812	0.605	0.007
KNN ***	pH	0.373	0.044	0.035	0.010
	MC	0.847	0.801	0.627	0.007
MLR	pH	0.190	0.049	0.038	0.011
	MC	0.857	0.777	0.580	0.007
PLSR	pH	0.231	0.047	0.038	0.011
	MC	0.864	0.759	0.574	0.007

* The layers and number of neurons for the ANN model are 11-10-70-2, the Epoch value is 500, and the activation function is *tanh-tanh* for each output parameter. ** Numbers of pH and moisture content splits are 2 and 5, respectively. *** The k value is 10 for both pH and moisture content.

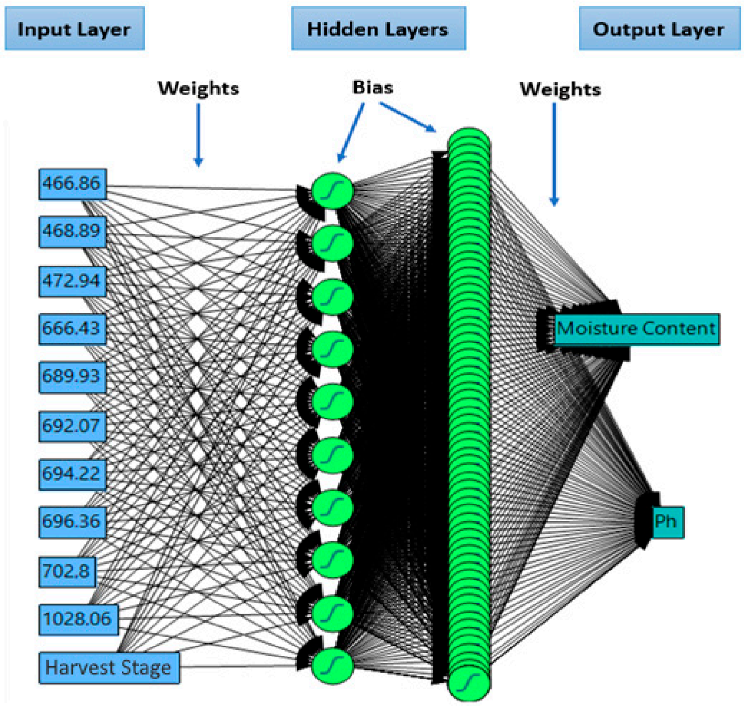
Predictor	Contribution	Portion	Rank
Harvest	45.6861	0.1668	1
694.22	18.3088	0.0668	2
689.93	16.9594	0.0619	3
696.36	16.9354	0.0618	4
692.07	14.1313	0.0516	5
702.8	9.8147	0.0358	6
666.43	6.8347	0.0250	7
662.17	6.4600	0.0236	8
683.51	5.5947	0.0204	9
687.79	5.4847	0.0200	10
552.95	5.3435	0.0195	11
524.04	5.0978	0.0186	12
501.47	4.8535	0.0177	13
679.23	3.9551	0.0144	14
499.43	3.8110	0.0139	15

(MC)

Predictor	Contribution	Portion	Rank
Harvest	0.032177	0.1892	1
694.22	0.003401	0.0200	2
692.07	0.003070	0.0181	3
466.86	0.002888	0.0170	4
468.89	0.002457	0.0145	5
472.94	0.001921	0.0113	6
1028.06	0.001865	0.0110	7
685.65	0.001844	0.0108	8
677.09	0.001809	0.0106	9
499.43	0.001796	0.0106	10
474.98	0.001765	0.0104	11
813.81	0.001587	0.0093	12
464.83	0.001574	0.0093	13
683.51	0.001439	0.0085	14
840.33	0.001370	0.0081	15

(pH)

(**a**)



(**b**)

Figure 2. Selected features (**a**) 15 best features for predicting MC and pH and (**b**) ANN structure for simultaneous detection of MC and pH using 11 optimum features.

The scatter plots representing the relationship between the actual and predicted values of each output parameter under various models are presented in Figures 3 and 4. In these scatter plots, the dark blue regions represent the 95% confidence intervals, estimating the range within which the mean difference between predicted and actual values is likely to fall. The light blue regions, on the other hand, represent the 95% prediction intervals. These intervals estimate the likely range of individual observations for specific predicted values, considering both the average difference and the natural variability of the observations. These confidence and prediction intervals help assess the reliability of the developed model's predictions and the degree of variability one might expect [46].

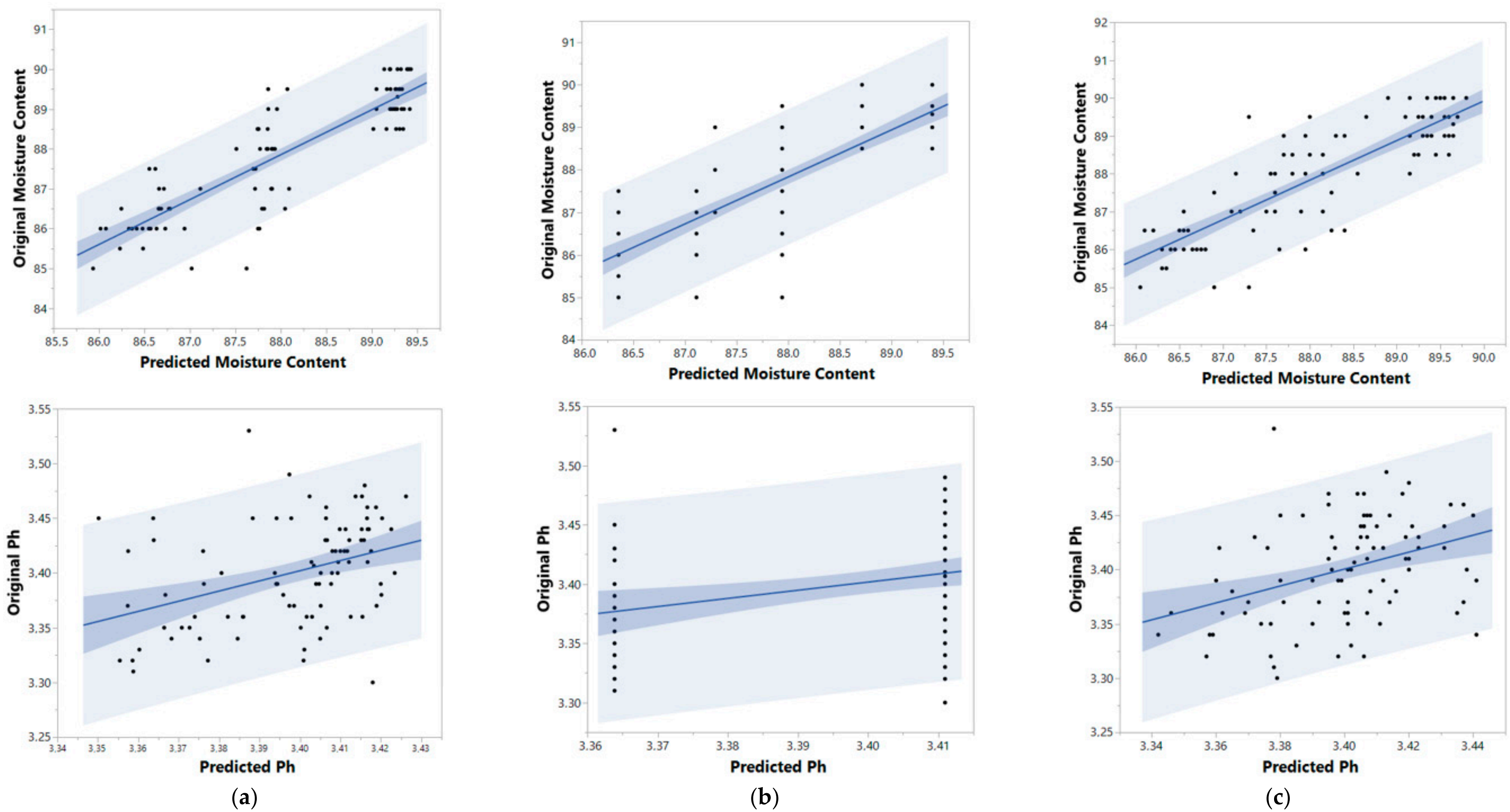


Figure 3. Scatter plot of the original and predicted values of target variables and the best linear regression line developed by (**a**) ANN; (**b**) DT; and (**c**) KNN models. The dark blue region depicts the 95% confidence interval, while the light blue region represents the 95% prediction interval.

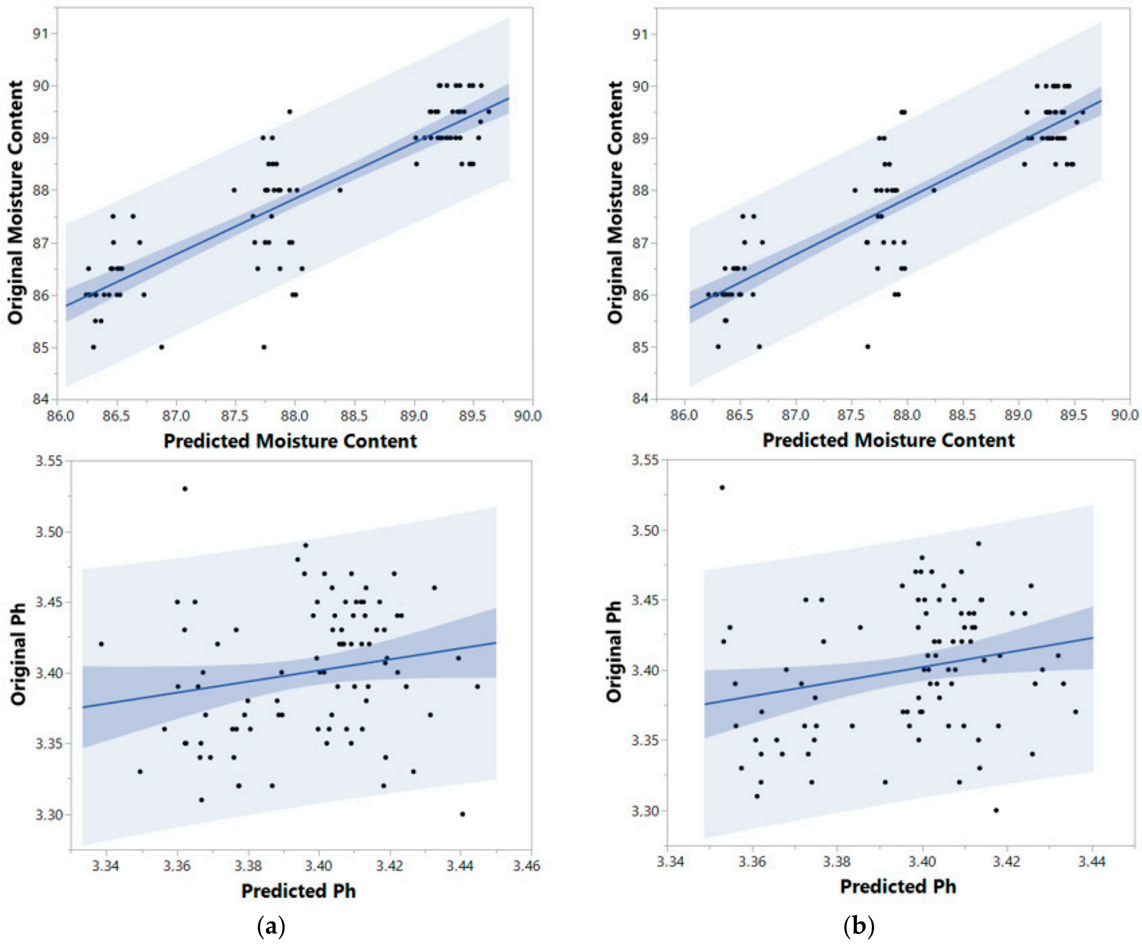


Figure 4. Scatter plot of the original and predicted values of target variables and the best linear regression line developed by (**a**) MLR and (**b**) PLSR models. The dark blue region depicts the 95% confidence interval, while the light blue region represents the 95% prediction interval.

It is worth mentioning the optimum MLR models as a function of selected features is provided below:

$$pH = 3.475 + (82.329 \times \text{“}466.86\text{”}) + (-119.253 \times \text{“}468.89\text{”}) + (36.684 \times \text{“}472.94\text{”}) + (0.526 \times \text{“}666.43\text{”}) + (27.218 \times \text{“}689.93\text{”}) + (-56.847 \times \text{“}692.07\text{”}) + (11.960 \times \text{“}694.22\text{”}) + (24.708 \times \text{“}696.36\text{”}) + (-6.780 \times \text{“}702.8\text{”}) + (-0.674 \times \text{“}1028.06\text{”}) + (0.006 \times Maturity)$$

$$MC = 90.062 + (-370.644 \times \text{“}466.86\text{”}) + (505.097 \times \text{“}468.89\text{”}) + (-135.043 \times \text{“}472.94\text{”}) + (-21.357 \times \text{“}666.43\text{”}) + (569.784 \times \text{“}689.93\text{”}) + (-1572.724 \times \text{“}692.07\text{”}) + (1444.348 \times \text{“}694.22\text{”}) + (-360.270 \times \text{“}696.36\text{”}) + (-66.760 \times \text{“}702.8\text{”}) + (7.371 \times \text{“}1028.06\text{”}) + (-1.181 \times Maturity)$$

4. Discussion

This study explored the capability of HSI imaging in the VIS-NIR range to predict apples' pH and MC in a non-destructive manner. Considering large data size is one of the main challenges in analyzing HSI data, data reduction techniques are recommended to capture the most informative trends for model development. Feature selection algorithms were utilized to select ten optimum wavelength features in addition to the maturity stage for developing calibration models. Statistical and machine learning techniques such as PLSR, MLR, kNN, DT, and ANN models were used to develop predictive models.

The maturity stage inherently influences the MC and pH of the apple, thus, it is expected to play a vital role in the developed predictive model. Moreover, our analysis highlighted the importance of five wavelengths (694.22, 689.93, 689.93, 696.36, and 692.07 nm) in predicting the apple MC. These wavelengths are associated with the functional groups that are important in determining water, sugar, and cellulose. The selected spectral bands for MC prediction were found to be concentrated between 666–702 nm, which is consistent with previous studies [13,47] that have identified these wavelengths as being informative for moisture content. Similarly, for pH prediction, the optimal wavelengths (694.22, 692.07, 466.86, 468.89, and 472.89 nm) were identified through feature selection. The spectral bands in this range are associated with functional groups such as carboxylic acids and amino acids that are important in determining pH. Specifically, the wavelengths between 464–499 nm and between 683–694 nm contained significant information for pH prediction. Overall, these results suggest that the selected features are closely related to the chemical composition of apples, and their inclusion in the predictive models can improve the accuracy of moisture content and pH predictions.

Among the models above, ANN provided the best performance with the prediction of MC and pH with *R*-values of 0.868 and 0.383, respectively. The obtained results have three main contributions in HSI-based apple quality monitoring: (1) ANN could outperform the commonly used PLSR predictive model in analyzing data, (2) VIS-NIR HSI is a promising non-destructive tool for predicting Pink Lady apples MC, but it lacks the desired accuracy for predicting the pH content, (3) integrating the findings of the present work with other relevant works, one can consider VIS-NIR HSI system as a single tool to predict apple firmness, SSC [33], phenolic content, antioxidant activity, ascorbic acid [4] and MC content.

Indeed, this study has highlighted some limitations in the prediction accuracy of pH values using the five applied machine-learning methods for the Pink Lady apples. The performance shortfall can be attributed to several factors, among which the relatively narrow data range observed for pH levels in the apple samples is a significant contributor. Moreover, the penetration depth of the utilized wavelength range into the apple tissue, specific to the studied variety, may have been limited. This likely resulted in a less comprehensive set of spectral data for accurate pH prediction, underscoring the impact of wavelength selection on the efficacy of HSI applications. In addition to these aspects, the suitability of the employed machine learning algorithms must be scrutinized. While these algorithms have shown effectiveness across a multitude of tasks, they may not be optimally tailored to predict pH levels derived from HSI data. It is plausible that more advanced machine learning models will better navigate the intricacy and variability of hyperspectral data, offering enhanced performance.

The results of other relevant works should be concisely discussed to compare the performance of apple MC and pH predictive models with those of other fruits. Rahman et al. [12] used NIR HSI (1000–1550 nm) and the PLSR model to investigate the non-destructive estimation of MC and pH in intact tomatoes. The *R*-values for MC and pH prediction were between 0.710 to 0.810 and 0.370 to 0.710, respectively. Although the fruit types were different, the higher reliability in predicting MC is in agreement with our observation for apples. Dong and Guo [17] utilized NIR HSI in the range of 900–1700 nm together with successive projection algorithm and least squares support vector machine to predict Fuji apple's MC and pH and reported *R*-values of 0.984 and 0.882, respectively. Their better performance compared to our results can be mainly attributed to the different spectral ranges they utilized. It is well-known that instruments working in the NIR range are more expensive than VIS-NIR as they utilize InGaAs instead of silicon detectors. Alternatively, different apple cultivars and modeling tools may also account for the discrepancy between the results. In a separate study, Wang et al. [48] utilized VIS-NIR transmittance spectroscopy to estimate pH in Fuji apples, achieving *R*-values ranging from 0.63 to 0.69. Once again, variations in apple types and different data-capturing techniques could potentially explain their superior performance. Overall, comparing our results with those of other relevant works highlights the potential of non-destructive techniques for predicting fruit quality parameters and the importance of considering various factors such as fruit type, spectral range, and modeling approaches when interpreting and comparing the results of different studies.

5. Future Work

This study provides a foundation for future exploration into enhancing the prediction accuracy of machine learning models using VIS-NIR HSI for apple quality assessment. It identifies numerous promising avenues for future research.

Broadening the sample size is a key opportunity for improving the models' robustness and generalizability. More comprehensive data could enhance the accuracy of the machine learning algorithms in recognizing and predicting patterns. Along with this, expanding the application of these predictive models to different apple varieties could be insightful, given the characteristic variation across cultivars. Further, advancements can be made in feature selection techniques. The exploration of methods such as genetic algorithms or principal component analysis might reveal more meaningful predictors within HSI data. The examination of state-of-the-art machine learning models and ensemble learning approaches, which employ multiple models for predictions, also promises potential improvements.

Another potential area for improvement lies in further fine-tuning the model hyperparameters. Optimization techniques, including grid search or random search, could be used to enhance predictive performance. Simultaneously, the wavelength range used in HSI can be expanded. A broader or different range of wavelengths could capture more spectral information, thereby better representing the properties of interest, such as pH levels. Advanced data augmentation techniques, like generative adversarial networks (GANs), also hold promise [49,50]. While traditional methods have served their purpose, GANs could enrich the dataset by generating synthetic yet realistic spectral data, proving particularly valuable in scenarios with limited original data or larger, more complex datasets. Another targeted focus for future research could be a holistic evaluation of various quality metrics of apples, such as MC, pH, firmness, SSC, starch, and acidity. In addition, to expedite the adoption of HSI technology within the fruit industry, it is crucial to conduct field tests that ascertain the practical performance of the developed technology.

6. Conclusions

The present work explored the feasibility of simultaneous detection of apple MC and pH using VIS-NIR HSI. Five different machine-learning models were employed to analyze HSI data. The correlation coefficients were identified as 0.868 and 0.383 for ANN, 0.844 and 0.307 for DT, 0.847 and 0.373 for KNN, 0.857 and 0.190 for MLR, and 0.864 and 0.231 for PLSR when predicting MC and pH, respectively. Hence, the predictions of the former and latter apple parameters were promising and poor, respectively. The superior performance of the ANN algorithms in analyzing VIS-NIR HSI data suggests that the well-known and commonly used PLSR approach is inferior to using ANN as a state-of-the-art intelligent algorithm. This exploration sets the stage for further research and developments in the field.

Author Contributions: E.K., visualization, software, implementing data curation, validation, AI and statistical methods, writing—original draft. N.Ç., conceptualization, formal analysis, investigation, methodology, data curation, resources, visualization, writing—original draft. B.Y., data curation, software, visualization, writing—original draft. M.N., results validation, writing—original, writing—revision. J.P., results validation, writing—revision. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding authors, upon reasonable request.

Conflicts of 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.

References

1. Faostat. Food and Agriculture Organization of the United Nations. Crops and Livestock Products. 2022. Available online: https://www.fao.org/faostat/en/#data/QCL (accessed on 14 March 2023).
2. Giovanelli, G.; Sinelli, N.; Beghi, R.; Guidetti, R.; Casiraghi, E. NIR spectroscopy for the optimization of postharvest apple management. *Postharvest Biol. Technol.* **2014**, *87*, 13–20. [CrossRef]
3. Tian, X.; Li, J.; Wang, Q.; Fan, S.; Huang, W. A bi-layer model for non-destructive prediction of soluble solids content in apple based on reflectance spectra and peel pigments. *Food Chem.* **2018**, *239*, 1055–1063. [CrossRef] [PubMed]
4. Çetin, N.; Sağlam, C. Rapid detection of total phenolics, antioxidant activity and ascorbic acid of dried apples by chemometric algorithms. *Food Biosci.* **2022**, *47*, 101670. [CrossRef]
5. Mo, C.; Kim, M.S.; Kim, G.; Lim, J.; Delwiche, S.R.; Chao, K.; Lee, H.; Cho, B.K. Spatial assessment of soluble solid contents on apple slices using hyperspectral imaging. *Biosyst. Eng.* **2017**, *159*, 10–21. [CrossRef]
6. Sağlam, C.; Çetin, N. Machine learning algorithms to estimate drying characteristics of apples slices dried with different methods. *J. Food Process. Preserv.* **2022**, *46*, e16496. [CrossRef]
7. Qin, J.; Lu, R. Measurement of the absorption and scattering properties of turbid liquid foods using hyperspectral imaging. *Appl. Spectrosc.* **2007**, *61*, 388–396. [CrossRef]
8. Lorente, D.; Aleixos, N.; Gómez-Sanchis, J.; Cubero, S.; García-Navarrete, O.L.; Blasco, J. Recent advances and applications of hyperspectral imaging for fruit and vegetable quality assessment. *Food Bioprocess Technol.* **2012**, *5*, 1121–1142. [CrossRef]
9. Lu, B.; Dao, P.D.; Liu, J.; He, Y.; Shang, J. Recent advances of hyperspectral imaging technology and applications in agriculture. *Remote Sens.* **2020**, *12*, 2659. [CrossRef]
10. ElMasry, G.; Wang, N.; El Sayed, A.; Ngadi, M. Hyperspectral imaging for non-destructive determination of some quality attributes for strawberry. *J. Food Eng.* **2007**, *81*, 98–107. [CrossRef]
11. Pu, Y.Y.; Sun, D.W. Vis–NIR hyperspectral imaging in visualizing moisture distribution of mango slices during microwave-vacuum drying. *Food Chem.* **2015**, *188*, 271–278. [CrossRef]
12. Rahman, A.; Kandpal, L.M.; Lohumi, S.; Kim, M.S.; Lee, H.; Mo, C.; Cho, B.K. Non-destructive estimation of moisture content, pH and soluble solid contents in intact tomatoes using hyperspectral imaging. *Appl. Sci.* **2017**, *7*, 109. [CrossRef]
13. Sun, Y.; Liu, Y.; Yu, H.; Xie, A.; Li, X.; Yin, Y.; Duan, X. 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. Methods* **2017**, *10*, 1535–1546. [CrossRef]
14. Lin, X.; Xu, J.L.; Sun, D.W. Investigation of moisture content uniformity of microwave-vacuum dried mushroom (*Agaricus bisporus*) by NIR hyperspectral imaging. *LWT* **2019**, *109*, 108–117. [CrossRef]
15. Cho, J.S.; Choi, J.Y.; Moon, K.D. Hyperspectral imaging technology for monitoring of moisture contents of dried persimmons during drying process. *Food Sci. Biotechnol.* **2020**, *29*, 1407–1412. [CrossRef] [PubMed]
16. Baiano, A.; Terracone, C.; Peri, G.; Romaniello, R. Application of hyperspectral imaging for prediction of physico-chemical and sensory characteristics of table grapes. *Comput. Electron. Agric.* **2012**, *87*, 142–151. [CrossRef]
17. Dong, J.; Guo, W. Non-destructive determination of apple internal qualities using near-infrared hyperspectral reflectance imaging. *Food Anal. Methods* **2015**, *8*, 2635–2646. [CrossRef]
18. Pu, H.; Liu, D.; Wang, L.; Sun, D.W. Soluble solids content and pH prediction and maturity discrimination of lychee fruits using visible and near infrared hyperspectral imaging. *Food Anal. Methods* **2016**, *9*, 235–244. [CrossRef]
19. Zhu, H.; Chu, B.; Fan, Y.; Tao, X.; Yin, W.; He, Y. Hyperspectral imaging for predicting the internal quality of kiwifruits based on variable selection algorithms and chemometric models. *Sci. Rep.* **2017**, *7*, 7845. [CrossRef]
20. Li, X.; Wei, Y.; Xu, J.; Feng, X.; Wu, F.; Zhou, R.; Jin, J.; Xu KYu, X.; He, Y. SSC and pH for sweet assessment and maturity classification of harvested cherry fruit based on NIR hyperspectral imaging technology. *Postharvest Biol. Technol.* **2018**, *143*, 112–118. [CrossRef]
21. Basak, J.K.; Madhavi BG, K.; Paudel, B.; Kim, N.E.; Kim, H.T. Prediction of Total Soluble Solids and pH of Strawberry Fruits Using RGB, HSV and HSL Colour Spaces and Machine Learning Models. *Foods* **2022**, *11*, 2086. [CrossRef]
22. Hosainpour, A.; Kheiralipour, K.; Nadimi, M.; Paliwal, J. Quality assessment of dried white mulberry (*Morus alba* L.) using machine vision. *Horticulturae* **2022**, *8*, 1011. [CrossRef]
23. Sabzi, S.; Nadimi, M.; Abbaspour-Gilandeh, Y.; Paliwal, J. Non-destructive estimation of physicochemical properties and detection of ripeness level of apples using machine vision. *Int. J. Fruit Sci.* **2022**, *22*, 628–645. [CrossRef]
24. Nadimi, M.; Divyanth, L.G.; Paliwal, J. Automated detection of mechanical damage in flaxseeds using radiographic imaging and machine learning. *Food Bioprocess Technol.* **2022**, *16*, 526–536. [CrossRef]
25. Benos, L.; Tagarakis, A.C.; Dolias, G.; Berruto, R.; Kateris, D.; Bochtis, D. Machine learning in agriculture: A comprehensive updated review. *Sensors* **2021**, *21*, 3758. [CrossRef]
26. Garavand, A.; Salehnasab, C.; Behmanesh, A.; Aslani, N.; Zadeh, A.H.; Ghaderzadeh, M. Efficient model for coronary artery disease diagnosis: A comparative study of several machine learning algorithms. *J. Healthc. Eng.* **2022**, *2022*, 5359540. [CrossRef] [PubMed]
27. Ghaderzadeh, M.; Asadi, F.; Hosseini, A.; Bashash, D.; Abolghasemi, H.; Roshanpour, A. Machine learning in detection and classification of leukemia using smear blood images: A systematic review. *Sci. Program.* **2021**, *2021*, 9933481. [CrossRef]

28. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine learning in agriculture: A review. *Sensors* **2018**, *18*, 2674. [CrossRef]
29. Pathan, M.; Patel, N.; Yagnik, H.; Shah, M. Artificial cognition for applications in smart agriculture: A comprehensive review. *Artif. Intell. Agric.* **2020**, *4*, 81–95. [CrossRef]
30. Shaikh, T.A.; Rasool, T.; Lone, F.R. Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. *Comput. Electron. Agric.* **2022**, *198*, 107119. [CrossRef]
31. Sharma, A.; Jain, A.; Gupta, P.; Chowdary, V. Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access* **2020**, *9*, 4843–4873. [CrossRef]
32. Wang, C.; Liu, B.; Liu, L.; Zhu, Y.; Hou, J.; Liu, P.; Li, X. A review of deep learning used in the hyperspectral image analysis for agriculture. *Artif. Intell. Rev.* **2021**, *54*, 5205–5253. [CrossRef]
33. Çetin, N.; Karaman, K.; Kavuncuoğlu, E.; Yıldırım, B.; Jahanbakhshi, A. Using hyperspectral imaging technology and machine learning algorithms for assessing internal quality parameters of apple fruits. *Chemom. Intell. Lab. Syst.* **2022**, *230*, 104650. [CrossRef]
34. Pourdarbani, R.; Sabzi, S.; Kalantari, D.; Paliwal, J.; Benmouna, B.; Martínez, J.M.M. Estimation of different ripening stages of Fuji apples using image processing and spectroscopy based on the majority voting method. *Comput. Electron. Agric.* **2020**, *176*, 105643. [CrossRef]
35. Pourdarbani, R.; Sabzi, S.; García-Mateos, G.; Paliwal, J.; Molina-Martínez, J.M. Using metaheuristic algorithms to improve the estimation of acidity in Fuji apples using NIR spectroscopy. *Ain Shams Eng. J.* **2022**, *13*, 101776. [CrossRef]
36. Yagcıoglu, A. *Drying Techniques of Agricultural Products*; Publication No: 536; Ege University, Faculty of Agriculture: Bornova, Türkiye, 1999. (In Turkish)
37. Stegmayer, G.; Milone, D.H.; Garran, S.; Burdyn, L. Automatic recognition of quarantine citrus diseases. *Expert Syst. Appl.* **2013**, *40*, 3512–3517. [CrossRef]
38. Drazin, S.; Montag, M. Decision tree analysis using weka. In *Machine Learning-Project II*; University of Miami: Coral Gables, FL, USA, 2012; pp. 1–3.
39. Xu, M.; Watanachaturaporn, P.; Varshney, P.K.; Arora, M.K. Decision tree regression for soft classification of remote sensing data. *Remote Sens. Environ.* **2005**, *97*, 322–336. [CrossRef]
40. Nettleton, D.F.; Orriols-Puig, A.; Fornells, A. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artif. Intell. Rev.* **2010**, *33*, 275–306. [CrossRef]
41. Romero, J.R.; Roncallo, P.F.; Akkiraju, P.C.; Ponzoni, I.; Echenique, V.C.; Carballido, J.A. Using classification algorithms for predicting durum wheat yield in the province of Buenos Aires. *Comput. Electron. Agric.* **2013**, *96*, 173–179. [CrossRef]
42. Kramer, O. K-Nearest Neighbors. In *Dimensionality Reduction with Unsupervised Nearest Neighbors. Intelligent Systems Reference Library*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 51. [CrossRef]
43. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009. [CrossRef]
44. Sun, X.; Li, H.; Yi, Y.; Hua, H.; Guan, Y.; Chen, C. Rapid detection and quantification of adulteration in Chinese hawthorn fruits powder by near-infrared spectroscopy combined with chemometrics. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2021**, *250*, 119346. [CrossRef]
45. Colton, T. *Statistics in Medicine*; Little Brown and Co.: New York, NY, USA, 1974; p. 179.
46. Meeker, W.Q.; Hahn, G.J.; Escobar, L.A. *Statistical İntervals: A Guide for Practitioners and Researchers*; John Wiley & Sons: Hoboken, NJ, USA, 2017; Volume 541.
47. Crichton, S.; Sturm, B.; Hurlbert, A. Moisture content measurement in dried apple produce through visible wavelength hyperspectral imaging. In Proceedings of the ASABE Annual International Meeting Sponsored by ASABE, New Orleans, LA, USA, 26–29 July 2015; ASABE Paper No. 152186400.
48. Wang, F.; Zhao, C.; Yang, H.; Jiang, H.; Li, L.; Yang, G. Non-destructive and in-site estimation of apple quality and maturity by hyperspectral imaging. *Comput. Electron. Agric.* **2022**, *195*, 106843. [CrossRef]
49. Lu, Y.; Chen, D.; Olaniyi, E.; Huang, Y. Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review. *Comput. Electron. Agric.* **2022**, *200*, 107208. [CrossRef]
50. Zhu, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Generative adversarial networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5046–5063. [CrossRef]