

Article

Hyperspectral Imaging Using a Convolutional Neural Network with Transformer for the Soluble Solid Content and pH Prediction of Cherry Tomatoes

Hengnian Qi ^{1,†} , Hongyang Li ^{1,†}, Liping Chen ^{2,*}, Fengnong Chen ³ , Jiahao Luo ¹ and Chu Zhang ^{1,*} 

¹ School of Information Engineering, Huzhou University, Huzhou 313000, China

² Huzhou Agricultural Science and Technology Development Center, Huzhou 313000, China

³ School of Artificial Intelligence, Hangzhou Dianzi University, Hangzhou 310018, China

* Correspondence: chenliping20042000@163.com (L.C.); chuzh@zjhu.edu.cn (C.Z.)

† These authors contributed equally to this work.

Abstract: Cherry tomatoes are cultivated worldwide and favored by consumers of different ages. The soluble solid content (SSC) and pH are two of the most important quality attributes of cherry tomatoes. The rapid and non-destructive measurement of the SSC and pH of cherry tomatoes is of great significance to their production and consumption. In this research, hyperspectral imaging combined with a convolutional neural network with Transformer (CNN-Transformer) was utilized to analyze the SSC and pH of cherry tomatoes. Conventional machine learning and deep learning models were established for the determination of the SSC and pH. The findings demonstrated that CNN-Transformer yielded outstanding results in predicting the SSC, with the coefficient of determination of calibration (R^2_C), validation (R^2_V), and prediction (R^2_P) for the SSC being 0.83, 0.87, and 0.83, respectively. Relatively worse results were obtained for the pH value prediction, with R^2_C , R^2_V , and R^2_P values of 0.74, 0.68, and 0.60, respectively. Furthermore, the visualization of the CNN-Transformer model revealed the wavelength weight distributions, indicating that the 1380–1650 nm range served as the characteristic band for the SSC, while the spectral range at 945–1280 nm was the characteristic band for pH. In conclusion, integrating spectral information features with the attention mechanism of Transformer through a convolutional neural network can enhance the accuracy of predicting the SSC and pH for cherry tomatoes.

Keywords: tomato quality; wavelength weight visualization; deep learning regression; non-destructive testing



Citation: Qi, H.; Li, H.; Chen, L.; Chen, F.; Luo, J.; Zhang, C. Hyperspectral Imaging Using a Convolutional Neural Network with Transformer for the Soluble Solid Content and pH Prediction of Cherry Tomatoes. *Foods* **2024**, *13*, 251. <https://doi.org/10.3390/foods13020251>

Academic Editors: Geraldine Dowling and Emilio Alvarez-Parrilla

Received: 17 November 2023

Revised: 19 December 2023

Accepted: 4 January 2024

Published: 12 January 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Cherry tomatoes (*Lycopersicon esculentum* var. *cerasiforme* A. Gray) are a popular tomato cultivar rich in lycopene, natural organic acids, essential human amino acids, vitamin C, and various mineral elements. Cherry tomatoes have become an indispensable item on the table with multiple health benefits, including the ability to warm the lung, liver, spleen, and stomach; quench thirst; and prevent cartilage disease and gum bleeding in children [1,2]. Quality indicators such as the soluble solid content (SSC) and pH are crucial for decision making in cherry tomato harvesting, trading, and consumption. However, the current quality judgment of cherry tomatoes primarily depends on the color, size, and shape of the fruits, leading to high labor costs and subjective judgments [3].

While using a refractometer for the measurement of the SSC and a pH meter for the measurement of pH can ensure high efficiency and accuracy in single-sample measurements, these methods require destructive sampling and cannot suffice large-scale measurement needs. Hyperspectral imaging technology comprises two-dimensional imaging and one-dimensional spectral technology, with several advantages such as a high resolution, multi-band imaging, and the efficient integration of images. It has been proven

as a valuable tool for advanced non-destructive testing in fields like plant disease detection, industrial food packaging, medical imaging classification, horticultural products [4–6]. It provides reliable, non-invasive, and quick analyses of essential nutrients like soluble solids, water content, and total acid in the quality detection of fruits [7–10].

Currently, modeling methods for fruit quality attribute prediction involve pre-processing wavelength-screening techniques, such as competitive adaptive reweighted sampling (CARS) and the successive projections algorithm (SPA) to enhance the features while reducing the dimension, and machine learning approaches such as partial least squares (PLS) and the support vector regression (SVR) algorithm are employed for modeling [11–13]. Rahman investigated the use of hyperspectral imaging to assess chemical components such as the moisture content (MC), pH, and SSC in intact tomatoes [14]. By developing multivariate corrected models for different pre-processed spectra using PLS regression, the study established that the regression model based on the first-order derivative of Savitzky–Golay (S-G) performed better for MC, pH, and SSC, with correlation coefficients of 0.81, 0.69, and 0.74, respectively. The root mean square errors of prediction were 0.63%, 0.06%, and 0.33%, respectively. Similarly, Sun explored the rapid non-destructive detection of cherry tomato sugar content in the field using the portable near-infrared spectroscopy instrument AMBER II [15]. The uninformative variable elimination–successive projections algorithm (UVE-SPA) was used to extract 12 characteristic wavelength points for PLS modeling, resulting in a coefficient of determination (R^2) of 0.9385 and 0.9347 for the training set and prediction set, respectively. The application of hyperspectral imaging (HSI) in the detection of tomato quality indicators can be divided into two categories: (1) For the detection of qualitative indicators (damage identification, spoilage detection, and maturity detection), Cho et al. [16] established the detection model of tomato maturity using the support vector machine (SVM) algorithm, and the classification accuracy reached more than 75.00%. (2) For the quantitative detection of the quality ingredients using HIS, Li Liu et al. [17] predicted the SSC of cherry tomatoes using PLS and LS-SVM based on HIS; Anisur Rahman et al. [14] determined the SSC, MC, and pH of cherry tomatoes using PLS based on HSI.

Deep learning enables models to automatically identify features through a vast range of sample data. It is widely used in various industries, including medicine, agriculture, object detection, image recognition, signal processing, and many other fields. By leveraging deep learning regression models and hyperspectral images, Xiang created non-destructive testing techniques for both SSC and fruit hardness [18]. Their findings highlight a new one-dimensional convolutional ResNet-based regression model. When compared to existing regression techniques, the proposed method obtained superior results that showed a 26.4% increase for SSC predictions and a 33.7% increase for hardness, particularly with larger sample sizes. Zhang utilized near-infrared hyperspectral imaging (NIR-HSI) to determine the total phenols, total flavonoids, and total anthocyanins present in black goji berries [19]. Furthermore, they developed a convolutional neural network (CNN) to predict the chemical composition, which yielded comparable performance to conventional methods such as PLS and the least-squares support vector machine (LS-SVM).

At present, there are still some technical bottlenecks to be overcome in the non-destructive testing of tomatoes using hyperspectral imaging technology, such as the improvement of spectral analysis algorithms, the upgrading of instruments and equipment, the control of detection costs, etc. The research and development of sensor hardware is relatively slow. Therefore, the current research of hyperspectral imaging technology in the field of tomato quality detection mainly focuses on data mining in order to improve the accuracy and efficiency of tomato quality detection. Classical modeling methods have the advantage of being easily comprehensible and implementable, and are known to yield positive results with small datasets. Nevertheless, they also exhibit certain disadvantages. For one, it necessitates human intervention to select features and algorithms, requiring operators with high levels of experience and expertise. Furthermore, it relies on the selection and processing of features, which can potentially impact the stability and reliability of

modeling outcomes. Deep learning approaches possess end-to-end modeling capabilities that can automatically learn features and patterns from raw data, thereby yielding effective classification and regression outcomes. This technique displays exceptional adaptability to data and generalization abilities, enabling it to handle very large datasets and highly nonlinear problems. In addition, deep learning models are characterized by their high scalability and flexibility, and can effectively process diverse types of data, including images, sound, text, and time-series data. In 2017, the Transformer was introduced as a neural network model capable of processing sequence data exclusively through the use of the attention mechanism [20]. Originally designed for natural language processing (NLP) tasks, Transformers have demonstrated remarkable flexibility and performance, gaining popularity in various applications, such as computer vision (CV) and audio processing. With the advantages in vision tasks, video processing, and music generation, the Transformer framework implemented in Vision Transformer (ViT) and BERT facilitates a performance enhancement of their models specifically tailored to the task at hand. Therefore, this study presented a Transformer-based regression model to predict the SSC and pH of cherry tomatoes using hyperspectral data. Although CNN approaches have been widely used for spectra data analysis due to their ability to learn deep features [19], the Transformer, as a different deep learning approach with a strong ability to learn deep features, has not been used for regression with spectra.

This study employed a combination of hyperspectral imaging techniques and deep learning-based regression models to predict the SSC and pH of cherry tomatoes, and verified the self-attention mechanism of the Transformer for regression using spectra data for the first time. The combination of a CNN and the Transformer was also explored to enhance the feature-learning ability of the two deep learning approaches. This study is guided by three specific aims: (1) to establish traditional regression models (PLSR, SVR) and deep learning regression models (CNN, LSTM, and Transformer) for SSC and pH prediction; (2) to improve the CNN-Transformer model regression effect; and (3) to promote model interpretability between diverse wavelengths and the SSC and pH.

2. Materials and Methods

2.1. Sample Preparation

In mid-July 2022, a total of 357 cherry tomatoes (cv. Zheyangfen 1) were collected in five batches from a greenhouse located in Changxing County, Huzhou, Zhejiang Province, China. These samples comprised 177 ripe tomatoes, with over 90% displaying a red fruit color, and 180 immature tomatoes, with over 70% displaying a red fruit color. In order to control the experimental variables, all samples were intact and the fruit surface was clean. The fruit pedicle was uniformly removed for data collection. Hyperspectral image acquisition and reference value measurement of the SSC and pH were accomplished on the day of sample collection of each batch of samples.

2.2. SSC and pH Measurements

The SSC content was tested using a sugar refractometer (BR0035, WIGGENS Technology Ltd., Beijing, China), while the pH was determined using a Sartorius pH meter (PB-10, Sartorius, Gottingen, Germany). Upon the completion of hyperspectral image acquisition, a solitary cherry tomato sample was juiced via manual juicing and subsequently filtered with gauze. The SSC content was ascertained by dispensing 1 mL of juice into the refractometer, whilst the pH was measured through the use of the pH meter probe within the juice. To prevent any potential contamination, the refractometer, pH meter probe, and juicer were carefully washed with distilled water after each measurement and wiped clean with paper towels prior to the next reading. Expressed in Table 1, the resultant sample measurements are featured.

Table 1. Results of sample SSC and pH measurements.

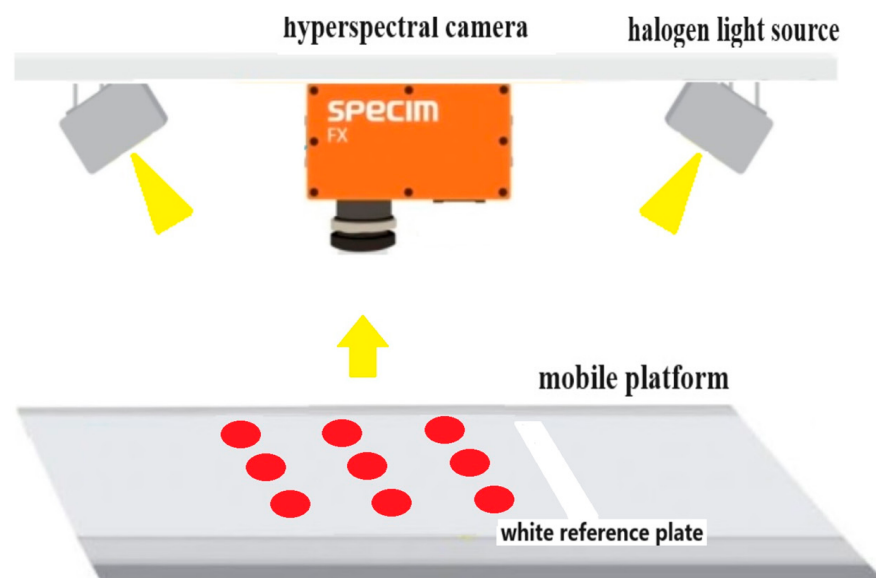
Number of Samples	SSC (°Brix)				pH			
	Mean	Maximum	Minimum	Standard Deviation	Mean	Maximum	Minimum	Standard Deviation
357	6.90	10.20	4.20	1.41	4.20	4.78	3.77	0.20

2.3. Hyperspectral Imaging System

In this study, hyperspectral images of cherry tomatoes were collected using a hyperspectral imaging system. The hyperspectral camera was FX-17 (Spectral Imaging Ltd., Oulu, Finland), the spectral range was 900–1700 nm, and the number of spectral bands was 224. The spectral resolution was 8 nm. Two groups of halogen light sources were symmetrically located on both sides of the hyperspectral camera for illumination, and each group of light sources consisted of three halogen bulbs. Each halogen bulb has a power of 35 w. During image acquisition, the moving speed of the mobile platform was set to 25 mm/s, the exposure time was set to 8.26 ms, and the hyperspectral image was obtained through push-scan imaging (line scanning) from left to right, as shown in Figure 1. The cherry tomatoes were evenly placed on the platform in 3 rows and 5 columns in a single capture process. Then, a white reference plate was placed on the front end to obtain the white reference image for light intensity calibration, and a black opaque lens cover with the lamp off was used to obtain the dark reference image [21]. The corrected image could be obtained by applying Formula (1). An RGB image and the grey-scale images at 1008, 1203, and 1403 nm of one hyperspectral image are shown in Figure S1 (in the Supplementary Materials).

$$R = \frac{I - D}{W - D} \quad (1)$$

where I is the original image, R is the corrected image, W is the white reference image, and D is the black reference image.

**Figure 1.** Hyperspectral imaging system.

2.4. Spectra Extraction and Processing

The entire region of each tomato in the hyperspectral images was defined as the region of interest (ROI). The ROIs were extracted using OpenCV image processing, and the average spectrum of all pixel directional spectra within the ROI was obtained to represent the spectral profile of the tomato. To ensure accurate detection results, the noisy ends

of the original spectral curve were eliminated by selecting the spectrum with smoother information. Box plots were employed to eliminate outliers, and multiplicative scatter correction (MSC) was utilized to pre-process the spectra. The process of spectrum extraction and processing is depicted in Figure 2.

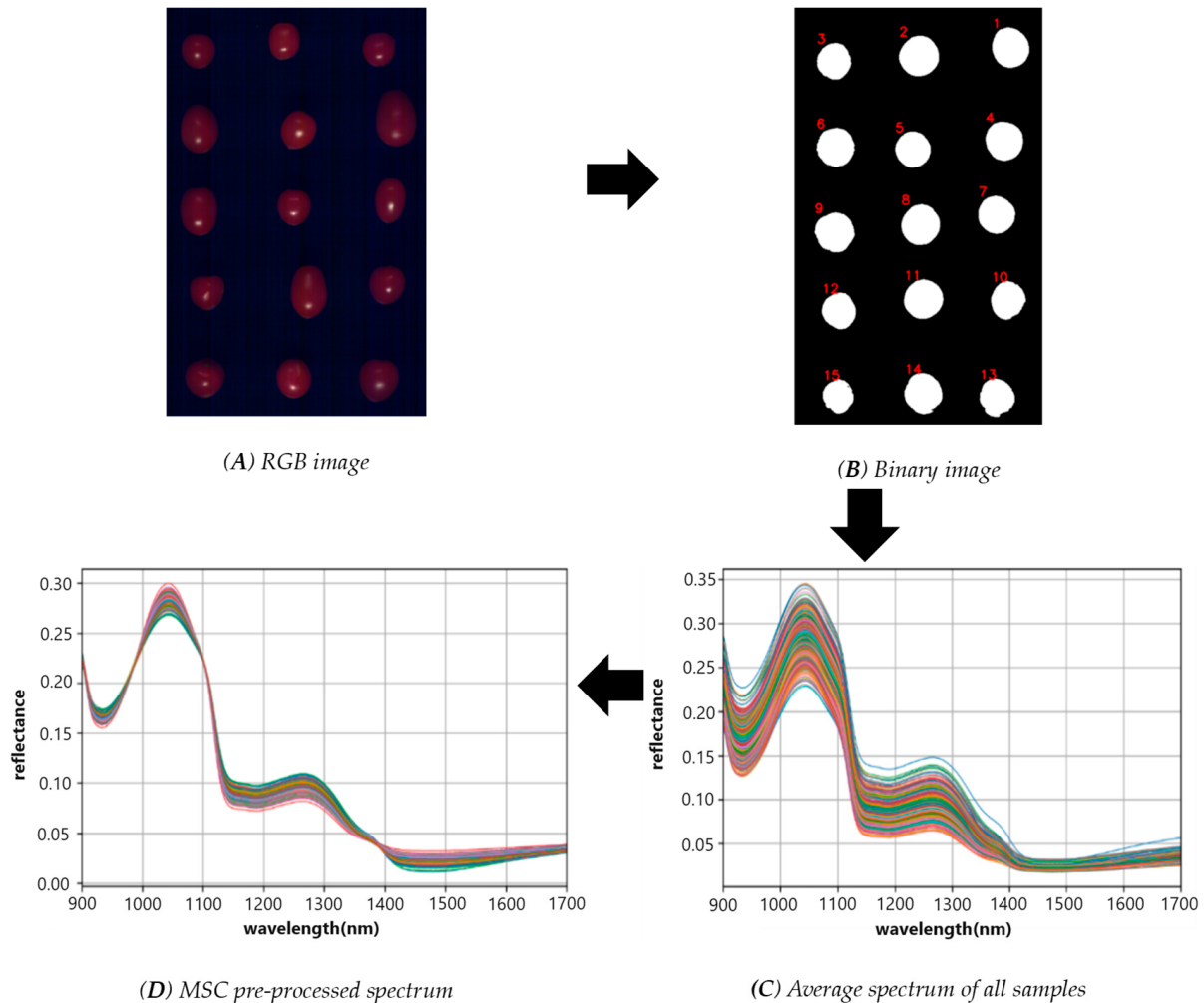


Figure 2. (A) RGB image; (B) the binary image of the cherry fruit region; (C) the spectra of all samples obtained by calculating the average of all the pixels within the same sample; and (D) the MSC pre-processed spectrum.

2.5. Partitioning the Sample Set

To establish and compare the predictive effects of different models, the samples were randomly divided into the calibration set, the validation set, and the prediction set at a ratio of 3:1:1. The calibration set was used to train the model, while the validation set was utilized to assess the model's robustness. The prediction set was used to evaluate the generalization capability of the final model. Since improper random data partitioning might result in regression overfitting, PLSR regression was employed to evaluate the performances of a random split of datasets. After trials, the datasets corresponding to the PLSR which had good prediction results and was free from overfitting were obtained and used for further analysis.

2.6. Regression Models

2.6.1. PLSR

Partial least squares regression (PLSR) [22] tackles the modeling problem involving an independent variable set (X) and dependent variable set (Y). Initially, the first pair of

principal components (t and u) are derived from X and Y, respectively, with the aim of maximizing their correlation degree. This process of extracting principal components is repeated until the regression equation reaches an acceptable level of accuracy. Ultimately, this approach can serve as a useful linear method in studying both spectrum and reference values. For this particular research, the number of principal components to be tuned was utilized in the range of 1–21. The optimal number of principal components was determined through the use of an automated parameter adjustment technique, known as GridSearchCV (function of model_selection in sklearn, grid search to find the global optimal value with cross-validation).

2.6.2. SVR

SVR is an application of support vector machines (SVMs) that presents unique advantages in addressing small-sample, nonlinear, and high-dimensional pattern recognition challenges [23,24]. In linear cases, an SVM constructs classifiers by maximizing the distance between different classifications and learning linear functions to classify samples. In non-linear cases, the SVM transforms low-dimensional features into high-dimensional feature spaces via kernel functions, causing problems to become linear in that space. Ultimately, an optimal classification hyperplane is established in the feature space, and the SVM model produces an optimal solution. This study uses various kernel functions such as linear, polynomial, and Gaussian, combined with a cross-verified mesh search to optimize parameter gamma in the range of $[10^{-4}, 10^5]$.

2.6.3. CNN

The CNN serves as one of the representative algorithms of deep learning and utilizes convolutional check variables to extract features. The convolution kernel parameter sharing in the hidden layer combined with the sparsity of inter-layer connections enable the CNN to efficiently lattice features with minimal computation, making it a useful tool in computer vision, natural language processing, and other fields [21]. In the current study, the proposed CNN model included two one-dimensional convolution layers, utilizing 128 and 256 channels, respectively, with each convolution kernel size set to 1×3 . To reduce the influence of gradient disappearance and improve the calculation speed, the activation function was selected as “ReLU”. The dropout layer was employed after the fully connected layer, and the probability was set to 0.3 to prevent the overfitting of the model. To expedite the convergence rate, the learning rate was set to 0.0005, the batch size was set to 10, and the AdamW optimizer was employed during training.

2.6.4. Long Short-Term Memory (LSTM) and CNN-LSTM

The LSTM network is a specialized RNN model that addresses long-term dependency issues through a unique structural design. Unlike other models that require expensive measures to retain early information, the LSTM’s default behavior is to store such data. This is achieved through the integration of three distinct gates: forgetting, memory, and output gates. These gates are responsible for managing the retention and transmission of information within the LSTM, as detailed in the literature [25]. In this study, a linear transformation was utilized to pre-process data before inputting a one-way LSTM model with 1024 input dimensions and 1024 hidden-layer dimensions.

Additionally, for the CNN-LSTM model in this study, CNN feature extraction was implemented before inputting the LSTM, and three-layer, one-dimensional convolution and a 3×3 convolution kernel were adopted. The number of channels in the convolution kernels was 128, 256, and 512, respectively. The one-way LSTM with 512 inputs and 256 hidden dimensions was used for the CNN-LSTM.

2.6.5. Transformer and CNN-Transformer

The Transformer system comprises two fundamental components—an encoder and a decoder. The Multi-Head-Attention in the encoder uses multiple heads to form

multiple subspaces, which allows the model to focus on different aspects of information and learn complex understanding requirements in the context through the layer stacking of attention. In this study, only the encoder component of the Transformer was utilized to isolate the attention information, finally feeding it to the full connection layer regression output [20,26–28]. The Multi-Head-Attention function of the system can be defined as a process that maps a query and a set of key–value pairs to produce an output, where all components, namely the query, key, value, and output, are represented as vectors. The output is obtained by calculating a weighted sum of the provided values, where each weight is derived from the compatibility function between the query and its corresponding key. The multi-head enables the system to efficiently concentrate on diverse representation subspaces across multiple locations. The use of an attentive head can help to overcome the potential of averaging that may result in the suppression of useful information.

The CNN-Transformer model diagram is depicted in Figure 3. The initial step involved feeding the input spectral data through the convolutional layer to extract features, with the number of convolutional cores being 128 and 256 after two convolutional operations. For positional encoding, positional functions were utilized to enable the model to concentrate on relative positional information, as shown in Formula (2). The encoder layer, based on the debugging outcome, comprised 6 layers, and each layer was constructed with Multi-Head-Attention, Feed Forward, and Add and Normalize. The residual standardized connection consisted of Add and Norm, with its computation depicted in Formula (3). Multi-Head-Attention was the essence of the Transformer, and four heads were employed to form multiple subspaces that enable the model to concentrate on different aspects of the information, estimating the correlation between different spectral bands and the output residual connection of the convolutional layer, as shown in Formula (4). Finally, the predicted results were obtained by the fully connected neural network (FC layer).

$$\begin{aligned} \text{PE}(\text{pos}, 2i) &= \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right) \\ \text{PE}(\text{pos}, 2i + 1) &= \cos\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right) \end{aligned} \quad (2)$$

where PE indicates the position code, pos indicates the position of the current character in the input letter, and d_{model} is the representation dimension.

$$\text{LayerNorm}(X + \text{MultiHeadAttention}(X))$$

$$\text{LayerNorm}(X + \text{FeedForward}(X)) \quad (3)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where Q , K , and V are all obtained through the transformation of input spectral variables, and d_k represents the variable dimension.

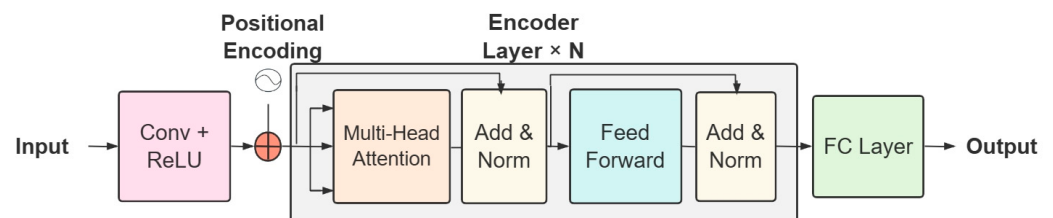


Figure 3. CNN-Transformer network structure schematic.

When training the CNN-Transformer model, the hyperparameters of the model were set, with a learning rate of 0.0008, a batch size of 10, a source sentence length of 209, a target length of 1, a word vector dimension of 256, a hidden layer in the twice linear layer of 1024, and a $K = Q = V$ dimension of 64. The number of encoder layers was set to 6; the number

of headers was set to 4. MSELoss was used as the loss function, and the AdamW optimizer was used for training to speed up the convergence.

2.7. Software, Hardware, and Performance Evaluation

All models were constructed using Python (version 3.6), with the deep learning framework PyTorch (version 1.9.0) utilized to establish the model's architecture and execute calculations on the GPU. The SVR and PLSR were modeled using Scikit-learn (version 0.24.2). All data analysis was carried out on a computer containing an Intel (R) Core (TM) i5-10500H processor, operating at 2.50 GHz, and 16 GB of RAM, alongside an NVIDIA GeForce RTX 3050 Laptop GPU.

To compare the performances of the PLSR, SVR, CNN, LSTM and Transformer models, the root mean square error (RMSE) (utilized as Formula (5)), the coefficient of determination (R^2) (utilized as Formula (6)), and the relative percentage difference (RPD) (utilized as Formula (7)) were used as performance evaluation indicators.

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2} \quad (5)$$

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y}_i - y_i)^2} \quad (6)$$

y_i : true value; $\hat{y}_i$: predicted value; $\bar{y}_i$: mean of true value.

$$\text{RPD} = \frac{SD}{SEP} \quad (7)$$

SD : the standard deviation; SEP : the root mean square error

The R^2 provides insight into the ability of the independent variable (x) to account for changes in the dependent variable (y). A higher R^2 value indicates a better fit for the model. Likewise, the RMSE gauges the mean square variance between predicted and actual result values for the model. A lower RMSE is indicative of a better model performance. $\text{RPD} < 1.4$ means that the model is not reliable, $1.4 < \text{RPD} < 2.0$ means that the model is considered reliable, and $\text{RPD} > 2.0$ means that the built model has high reliability and can be used for model analysis.

3. Results and Discussion

3.1. Spectral Profiles

In this study, 357 cherry tomato samples were analyzed and studied. The wavelength range of the full spectrum data was 900–1700 nm, and the noise at both ends of the original spectral curve was relatively large. The spectral bands between 900 and 945 nm were removed from the beginning, and only the spectra in the range of 946–1700 nm were used for further analysis. MSC was used for spectral correction. The principle is to conduct unary linear regression between the spectrum of each sample and the average spectrum, find the baseline translation and offset of each sample, and then correct the spectrum to eliminate the spectral difference caused by different scattering levels. Figure 2C shows the spectral data after MSC treatment; subtle differences can be observed in different samples, and two distinct peaks can be observed at 1000–1100 nm and 1220–1300 nm. The relationship between the spectrum and the SSC and pH was further discussed by means of machine learning and deep learning models.

3.2. Regression Model Establishment

To establish the regression models, the full-range spectra were used. The conventional PLSR and SVR methods were used to develop regression models using full-range spectra. Impressively, promising outcomes are observed and evidenced in Table 2. With regard to the SSC estimation, both models attained R_c^2 , R_v^2 , and R_p^2 values above 0.8; both models

attained RPDP values above 2.4. However, the PLSR model outperformed the SVR model in the pH prediction domain, with R_c^2 , R_v^2 , and R_p^2 values of approximately 0.6. PLSR obtained an RPDP value of 1.82, which was the best performance of all models on the prediction set. Anisur Rahman et al. used the spectra in the range of 1000–1550 nm, and after MSC treatment, established a PLS model to predict the SSC and pH, with an Rpred of 0.78 and 0.32, respectively [14]. Hence, it could be deduced from these findings that the combination of hyperspectral imaging with traditional regression methods showed an excellent outcome in SSC prediction, but its potential in pH prediction remains unveiled.

Table 2. Results of PLSR, SVR, CNN, LSTM, Transformer, CNN-LSTM, and CNN-Transformer models for SSC and pH prediction.

Indicators	Models	Calibration		Validation		Prediction		
		R_c^2	RMSEC	R_v^2	RMSEV	R_p^2	RMSEP	RPDP *
SSC	PLSR	0.80	0.55	0.82	0.58	0.84	0.56	2.44
	SVR	0.90	0.39	0.87	0.52	0.85	0.59	2.34
	CNN	0.50	0.95	0.68	0.81	0.50	0.97	1.41
	LSTM	0.56	0.89	0.54	0.98	0.56	0.91	1.52
	Transformer	0.62	0.82	0.73	0.75	0.72	0.73	1.88
	CNN-LSTM	0.57	0.88	0.55	0.98	0.56	0.91	1.50
	CNN-Transformer	0.83	0.58	0.87	0.52	0.83	0.56	2.45
pH	PLSR	0.66	0.10	0.61	0.12	0.73	0.10	1.82
	SVR	0.75	0.08	0.55	0.13	0.58	0.13	1.53
	CNN	0.60	0.12	0.58	0.12	0.54	0.13	1.48
	LSTM	0.47	0.14	0.42	0.15	0.48	0.14	1.39
	Transformer	0.51	0.14	0.56	0.13	0.63	0.12	1.64
	CNN-LSTM	0.73	0.10	0.53	0.13	0.61	0.12	1.60
	CNN-Transformer	0.74	0.10	0.69	0.11	0.60	0.12	1.59

* RPDP: The difference in precision of RMSE results in the difference in RPD.

Additionally, deep learning methods including CNN, Transformer, and LSTM were used for regression analysis. Using the feature extraction ability of the CNN, the model parameters were adjusted to build the CNN-LSTM and CNN-Transformer models. In the prediction of the SSC, the R_c^2 , R_v^2 , and R_p^2 values of the CNN and LSTM models were found to be between 0.5 and 0.6, and both models attained RPDP values above 1.5. The Transformer model performed relatively better, with an R^2 value around 0.7 (Table 2), and the RPDP value was 1.88. By incorporating spectral data features extracted by the CNN into the Transformer encoder, attention could be weighted to produce a more accurate prediction. It could be observed that the performance of the CNN-LSTM models were slightly improved compared to LSTM. The CNN-Transformer showed a particularly significant improvement compared to the Transformer, achieving R_c^2 , R_v^2 , and R_p^2 values of over 0.83 in SSC prediction. The scatter diagrams of the predicted and measured values of cherry tomatoes based on the CNN-Transformer model is shown in the Figure 4. The red line is the regression line representing the ideal correlation between the predicted value and the measured value, and the closer to the regression line, the better the model prediction is shown. It is worth mentioning the RPDP value of 2.45, which was the best performance of all models on the prediction set. In the prediction of pH, an R^2 near 0.69 and an RPDP near 1.6 (Figure 4) were obtained. In a similar study, Yun Xiang et al. used deep learning and hyperspectral images to predict the SSC and firmness of tomatoes. In terms of the SSC, the R^2 of the deep learning model was 0.49 in a small sample set (50 samples). Notably, the R^2 was 0.9 in a large sample (200 samples) of the deep learning model, and the result was better than PLSR, AdaBoost, KNNR, and SVR [18].

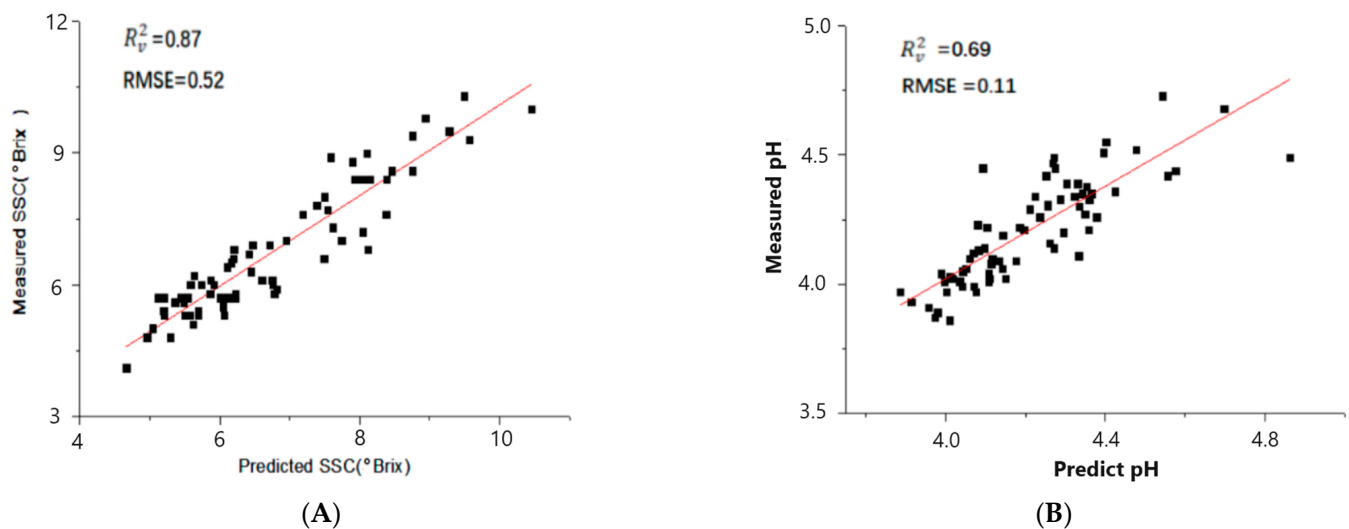


Figure 4. CNN-Transformer scatter diagrams. (A) Scatter diagram for SSC prediction validation set; (B) scatter diagram of pH prediction validation set.

In comparison, it was observed that the performances of SSC prediction exceeded the performances of pH prediction. However, it was worth emphasizing that the optimization of the deep learning model's structure could greatly improve its predictive capabilities. This was particularly true for the CNN-Transformer model, which exhibited remarkable performance in accurately predicting both the SSC and pH. Notably, the results showed that the SSC prediction accuracy obtained by the CNN-Transformer was comparable to that obtained by traditional methods such as PLSR and SVR. Moreover, it showcased excellent performance in predicting pH values. Empirical evidence unequivocally indicated the efficacy of the CNN-Transformer model in achieving precise and timely cherry tomato internal quality prediction. Additionally, the flexible structure of this model presents ample opportunities for further refinement and advancement, thereby augmenting its overall potential.

3.3. CNN-Transformer Network Visualization

The interpretability of the CNN-Transformer model was investigated by analyzing the weight distribution corresponding to different wavelengths in the prediction of the SSC and pH, respectively. Through the use of Grad-CAM, gradient information was extracted from the specified feature layer via backpropagation. This information was then utilized to calculate the importance of each channel within the feature layer, resulting in a heat map that allowed for an analysis of the model's acquisition of relevant feature information. While Grad-CAM was commonly employed for visualizing image classification problems, it could also be adapted to regression problems by utilizing gradient computation to capture the significance of input variables.

To explore interpretability, a randomly selected set of 50 samples was inputted into the pre-trained model. Grad-CAM was subsequently applied to extract gradient information from the final convolutional layer's feature map. The resulting weight vector and the associated channel of the feature map were linearly weighted and averaged to generate a one-dimensional weight distribution map. The weights corresponding to different wavelengths are shown in the Figure 5, where the blue shadow represents the variance of the samples, and the larger the shadow area, the larger the variance, indicating the greater the dispersion from the mean value.

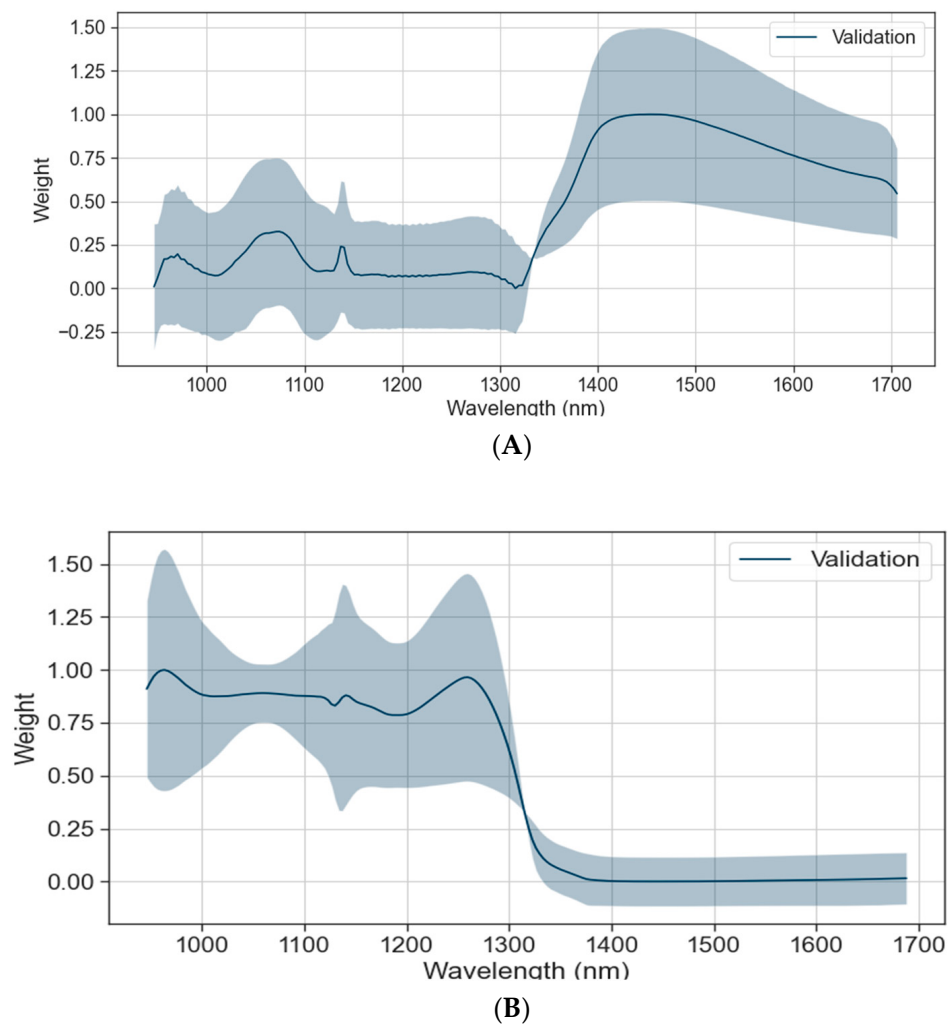


Figure 5. (A) Weight distribution corresponding to wavelength of CNN-Transformer calculated by Grad-CAM for SSC; (B) weight distribution corresponding to wavelength of CNN-Transformer calculated by Grad-CAM for pH.

3.3.1. Important Wavelengths for SSC Prediction

The Grad-CAM analysis showed that in the prediction of the SSC in the CNN-Transformer (Figure 5A), the weights of wavelengths in the range of 1380–1650 nm were the highest. The 1380 nm region was attributed to the influence of hydroxide in water, and it is speculated that the nearby wavelength is the main factor affecting the SSC. The weights of other wavelengths were relatively low, which also had an impact on the prediction performances.

Liu tested the sugar content of cherry tomatoes based on HSI [17,29]. Spectral curves between the wavelengths of 914.91 nm and 1661.91 nm were used to extract feature bands by CARS, SPA, and SPA-CARS. There were 36 wavelengths extracted by CARS, and there were 10 characteristic bands in the spectral range of 1380–1650 nm. There were 40 wavelengths extracted by SPA, and there were 24 characteristic bands in the spectral range of 1380–1650 nm. There were 35 wavelengths extracted by SPA, and there were 20 characteristic bands in the spectral range of 1380–1650 nm. The results were similar to the conclusion of this study. Li studied the online quality detection system of cherry tomatoes based on short-wave near-infrared spectroscopy. Nineteen variables were extracted from 900 to 1700 nm by CARS-PLS, nine variables were distributed from 900 to 1200 nm, and 10 variables were distributed from 1400 to 1700 nm, which was similar to the conclusion of this study [30,31]. The experimental results showed that the characteristic wavelength

range of 1380–1650 nm was the most important factor in predicting the SSC of cherry tomatoes, followed by the characteristic wavelength range of 900–1100 nm.

3.3.2. Important Wavelengths for pH Prediction

Based on the analysis conducted, it was shown that the wavelength range of 945–1280 nm held the highest significance in predicting the pH levels in the CNN-Transformer (Figure 5B). Li used the SPA and PLSR to predict the SSC and pH of cherry tomatoes in the 874–1734 nm region, and the distribution of pH characteristic bands in the spectral curve was highly consistent with the experimental results [9]. The range of wavelengths from 972 nm to 1200 nm appeared to have a significant impact on the prediction results. This prominence in prediction could be attributed to the vibration of hydrogen groups, especially the O-H and N-H groups. These hydrogen groups were commonly found in acidic compounds and were known to influence pH predictions.

3.4. Discussion

This study examined the impact of traditional methods and deep learning on predicting SSC and pH indicators of cherry tomatoes, with a focus on the Transformer model and CNN-Transformer model. The PLSR, SVR, and CNN-Transformer models all demonstrated excellent predictive performances for the SSC, with R^2_C , R^2_V , and R^2_P values all exceeding 0.8. PLSR and the CNN-Transformer also performed well in predicting pH, with R^2_C , R^2_V , and R^2_P values all exceeding 0.6. Research indicates that traditional methods such as PLSR and SVR remain reliable. However, by combining the extraction features of the CNN with the self-attention mechanism of the Transformer, it could be observed that the enhanced CNN-Transformer model has made significant progress in predicting SSC and pH in cherry tomatoes. This study observes that deep learning surpasses traditional machine learning algorithms in terms of better generalization across various metrics and adaptability to different datasets. Furthermore, in the context of predicting the levels of dry black goji berries, as well as flavonoids and anthocyanins in black wolfberry, the utilization of convolutional neural networks (CNNs) as a modeling and feature extraction method demonstrated favorable and comparable results when compared to the conventional approach [32]. The success of deep learning techniques indicates their promising potential as a valuable method for prediction and feature extraction in relevant fields.

The prediction effect of the pH of all models in this experiment was worse than that of the SSC, and the same conclusion could be found in other studies. Anisur Rahman et al. developed a multivariable calibration model using hyperspectral imaging and PLSR for pH and SSC, and the predictive effect of the pH was inferior to that of the SSC, with correlation coefficients of prediction (r_{pred}) of 0.75, 0.69, and 0.74 [14]. Yuping Huang et al. developed spatially resolved spectroscopy (SRS) to assess tomato quality. The prediction effect of the pH was lower than that of the SSC [33]. Hongbin Pu et al. employed hyperspectral imaging to forecast the sugar content and acidity of litchi. It could be found that the prediction of pH was worse than the prediction of SSC [34].

Grad-CAM was utilized for extracting gradient information from the specified feature layer through backpropagation to analyze the weight distribution corresponding to different wavelengths when predicting SSC and pH in the CNN-Transformer model. It was observed that the weight for predicting the SSC and pH differed significantly. The wavelength range of 1380–1650 nm was the key region for predicting the SSC, while the wavelength range of 945–1280 nm was the key region for predicting the pH. These findings align with those of other papers and demonstrate that the CNN-Transformer model could effectively learn spectral information for reliable prediction [9,17,29,31].

4. Conclusions

In this study, multiple predictive regression models such as PLSR, SVR, CNN, LSTM, CNN-LSTM, Transformer, and CNN-Transformer were developed to perform non-destructive testing of the SSC and pH of cherry tomatoes. The results showed that traditional machine

learning still has a stable performance in predicting SSC, but the effect of predicting pH is not so good. Deep learning methods have experienced great changes, among which the CNN-Transformer model has an excellent effect in predicting the SSC and pH value of cherry tomatoes; the R^2 values were around 0.87 and 0.7, respectively. The characteristic wavelength distribution of the CNN-Transformer model was visualized using Grad-CAM, which enhances the interpretability of the model. In conclusion, the CNN-Transformer could enhance the accuracy of predicting the SSC and pH of cherry tomatoes using hyperspectral imaging. It offers an efficient method for the non-destructive testing of the tomato quality index and other types of fruits. The application of hyperspectral imaging technology with deep learning in fruit quality detection can improve the quality and value of fruit. It can also drive and promote the development of the entire fruit industry chain and improve the competitiveness of the fruit industry by using advanced detection techniques and artificial intelligence.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/foods13020251/s1>: Figure S1: (a) RGB image of cherry tomato; (b1) corrected grayscale image at 1008 nm; (b2) corrected grayscale image at 1203 nm; (b3) corrected grayscale image at 1403 nm.

Author Contributions: Conceptualization, H.Q.; data curation, H.L.; formal analysis, H.L.; funding acquisition, H.L.; investigation, H.L. and J.L.; methodology, H.L. and C.Z.; project administration, H.Q. and L.C.; resources, L.C.; supervision, C.Z.; validation, F.C. and J.L.; visualization, F.C.; writing—original draft, H.Q. and H.L.; writing—review and editing, H.Q., H.L., L.C., F.C. and C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Postgraduate Research and Innovation Project of Huzhou University (grant number: 2023KYCX44), Huzhou Key R&D Program (grant number: 2023ZD2030) and the Zhejiang Province Key Laboratory of Smart Management and Application of Modern Agricultural Resources (grant number: 2020E10017).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data can be requested by contacting the corresponding authors via E-mail.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kumar, K.S.; Paswan, S.; Srivastava, S. Tomato—a natural medicine and its health benefits. *J. Pharmacogn. Phytochem.* **2012**, *1*, 33–43. [\[CrossRef\]](#)
2. Li, J.; Xiang, Z.; Wang, X.; Guo, Y.M.; Huang, Z.; Liu, L.; Li, X.; Du, Y. Current situation and prospect of tomato industry in China during the 13th Five-Year Plan. *Chin. Veg.* **2021**, 13–20. [\[CrossRef\]](#)
3. Wang, T.; Chen, J.; Fan, Y.; Qiu, Z.; He, Y. SeeFruits: Design and evaluation of a cloud-based ultra-portable NIRS system for sweet cherry quality detection. *Comput. Electron. Agric.* **2018**, *152*, 302–313. [\[CrossRef\]](#)
4. Huang, F.-H.; Liu, Y.-H.; Sun, X.; Yang, H. Quality inspection of nectarine based on hyperspectral imaging technology. *Syst. Sci. Control Eng.* **2021**, *9*, 350–357. [\[CrossRef\]](#)
5. Medus, L.D.; Saban, M.; Frances-Villora, J.V.; Bataller-Mompean, M.; Rosado-Muñoz, A. Hyperspectral image classification using CNN: Application to industrial food packaging. *Food Control* **2021**, *125*, 107962. [\[CrossRef\]](#)
6. Jeyaraj, P.R.; Samuel Nadar, E.R. Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. *J. Cancer Res. Clin. Oncol.* **2019**, *145*, 829–837. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Chandrasekaran, I.; Panigrahi, S.S.; Ravikanth, L.; Singh, C.B. Potential of Near-Infrared (NIR) Spectroscopy and Hyperspectral Imaging for Quality and Safety Assessment of Fruits: An Overview. *Food Anal. Methods* **2019**, *12*, 2438–2458. [\[CrossRef\]](#)
8. Leiva-Valenzuela, G.A.; Lu, R.; Aguilera, J.M. Prediction of firmness and soluble solids content of blueberries using hyperspectral reflectance imaging. *J. Food Eng.* **2013**, *115*, 91–98. [\[CrossRef\]](#)
9. Li, X.; Wei, Y.; Xu, J.; Feng, X.; Wu, F.; Zhou, R.; Jin, J.; Xu, K.; Yu, 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. [\[CrossRef\]](#)
10. Ma, T.; Xia, Y.; Inagaki, T.; Tsuchikawa, S. Rapid and nondestructive evaluation of soluble solids content (SSC) and firmness in apple using Vis–NIR spatially resolved spectroscopy. *Postharvest Biol. Technol.* **2021**, *173*, 111417. [\[CrossRef\]](#)

11. Mishra, P.; Marini, F.; Brouwer, B.; Roger, J.M.; Biancolillo, A.; Woltering, E.; Echte, E.H. Sequential fusion of information from two portable spectrometers for improved prediction of moisture and soluble solids content in pear fruit. *Talanta* **2021**, *223*, 121733. [CrossRef]
12. Gao, Q.; Wang, M.L.; Guo, Y.Y.; Zhao, X.Q.; He, D.J. Comparative Analysis of Non-Destructive Prediction Model of Soluble Solids Content for *Malus micromalus* Makino Based on Near-Infrared Spectroscopy. *IEEE Access* **2019**, *7*, 128064–128075. [CrossRef]
13. Sohrabi, M.M.; Ahmadi, E.; Monavar, H.M. Nondestructive analysis of packaged grape tomatoes quality using PCA and PLS regression by means of fiber optic spectroscopy during storage. *J. Food Meas. Charact.* **2017**, *12*, 949–966. [CrossRef]
14. Rahman, A.; Kandpal, L.M.; Lohumi, S.; Kim, M.S.; Lee, H.; Mo, C.; Cho, B.-K. Nondestructive estimation of moisture content, pH and soluble solid contents in intact tomatoes using hyperspectral imaging. *Appl. Sci.* **2017**, *7*, 109. [CrossRef]
15. Sun, Y.; Liu, C.; Sun, X.; Wen, S. Rapid sugar analysis model of cherry tomato based on portable near infrared spectrometer. *Food Ferment. Ind.* **2021**, *47*, 214–220. [CrossRef]
16. Cho, B.-H.; Kim, Y.-H.; Lee, K.-B.; Hong, Y.-K.; Kim, K.-C. Potential of Snapshot-Type Hyperspectral Imagery Using Support Vector Classifier for the Classification of Tomatoes Maturity. *Sensors* **2022**, *22*, 4378. [CrossRef]
17. Liu, L.; Shi, X.; Zhang, S.; Shi, Y.; Long, Y. Saccharinity Test on Cherry Tomatoes Based on Hyperspectral Imaging. *Int. J. Des. Nat. Ecodynamics* **2020**, *15*, 103–111. [CrossRef]
18. Xiang, Y.; Chen, Q.; Su, Z.; Zhang, L.; Chen, Z.; Zhou, G.; Yao, Z.; Xuan, Q.; Cheng, Y. Deep learning and hyperspectral images based tomato soluble solids content and firmness estimation. *Front. Plant Sci.* **2022**, *13*, 860656. Available online: <https://arxiv.org/pdf/2203.05199.pdf> (accessed on 3 January 2024). [CrossRef]
19. Zhang, B.; Zhang, M.; Shen, M.; Li, H.; Zhang, Z.; Zhang, H.; Zhou, Z.; Ren, X.; Ding, Y.; Xing, L.; et al. Quality monitoring method for apples of different maturity under long-term cold storage. *Infrared Phys. Technol.* **2021**, *112*, 103580. [CrossRef]
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [CrossRef]
21. Jin, B.; Qi, H.; Jia, L.; Tang, Q.; Gao, L.; Li, Z.; Zhao, G. Determination of viability and vigor of naturally-aged rice seeds using hyperspectral imaging with machine learning. *Infrared Phys. Technol.* **2022**, *122*, 104097. [CrossRef]
22. Cheng, J.-H.; Sun, D.-W. Partial least squares regression (PLSR) applied to NIR and HSI spectral data modeling to predict chemical properties of fish muscle. *Food Eng. Rev.* **2017**, *9*, 36–49. [CrossRef]
23. Malegori, C.; Nascimento Marques, E.J.; de Freitas, S.T.; Pimentel, M.F.; Pasquini, C.; Casiraghi, E. Comparing the analytical performances of Micro-NIR and FT-NIR spectrometers in the evaluation of acerola fruit quality, using PLS and SVM regression algorithms. *Talanta* **2017**, *165*, 112–116. [CrossRef]
24. Sarkar, S.; Basak, J.K.; Moon, B.E.; Kim, H.T. A Comparative Study of PLSR and SVM-R with Various Preprocessing Techniques for the Quantitative Determination of Soluble Solids Content of Hardy Kiwi Fruit by a Portable Vis/NIR Spectrometer. *Foods* **2020**, *9*, 1078. [CrossRef]
25. Hochreiter, S.; Schmidhuber, J. LSTM can solve hard long time lag problems. *Adv. Neural Inf. Process. Syst.* **1996**, *9*.
26. Clark, K.; Khandelwal, U.; Levy, O.; Manning, C.D. What does bert look at? an analysis of bert's attention. *arXiv* **2019**, arXiv:1906.04341.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
28. Lee, J.; Toutanova, K. Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
29. Liu, Y.; Zhang, Y.; Jiang, X.; Liu, H. Detection of the quality of juicy peach during storage by visible/near infrared spectroscopy. *Vib. Spectrosc.* **2020**, *111*, 103152. [CrossRef]
30. Li, P.; Li, S.; Du, G.; Jiang, L.; Liu, X.; Ding, S.; Shan, Y. A simple and nondestructive approach for the analysis of soluble solid content in citrus by using portable visible to near-infrared spectroscopy. *Food Sci. Nutr.* **2020**, *8*, 2543–2552. [CrossRef] [PubMed]
31. Li, L.; Peng, Y.; Yang, C.; Li, Y. Optical sensing system for detection of the internal and external quality attributes of apples. *Postharvest Biol. Technol.* **2020**, *162*, 111101. [CrossRef]
32. Zhang, C.; Wu, W.; Zhou, L.; Cheng, H.; Ye, X.; He, Y. Developing deep learning based regression approaches for determination of chemical compositions in dry black goji berries (*Lycium ruthenicum* Murr.) using near-infrared hyperspectral imaging. *Food Chem.* **2020**, *319*, 126536. [CrossRef]
33. Huang, Y.; Lu, R.; Chen, K. Assessment of tomato soluble solids content and pH by spatially-resolved and conventional Vis/NIR spectroscopy. *J. Food Eng.* **2018**, *236*, 19–28. [CrossRef]
34. 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]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.