An integrated learning algorithm for early prediction of melon harvest

Chunyang Qian, Taihang Du, Shuguang Sun, Wei Liu, Haiguang Zheng & Jianchun Wang

Different modeling techniques must be applied to manage production and statistical estimation to predict the expected harvest. By calculating advanced production methods and the rational valuation of different factors, we can accurately capture the variety of growth characteristics and the expected yield. This paper obtained 32 feature variables related to melons, including phenological features, shape features, and color features. The Gradient Boosted Decision Tree (GBDT) network and the Grid Search (GS) hyperparameter seeking method was applied to calculate the degree of importance of all melon fruits' characteristics and construct prediction models for three expected harvest indexes of melon yield, sugar content, and endocarp hardness. To facilitate growers to carry out prediction and estimation in the field without destroying the melon fruits. The reduced feature variables were selected as inputs. The GBDT model was used to provide a significant advantage in prediction compared to both Random Forest (RF) and Support Vector Regression (SVR) methods. In addition, to verify the feasibility of using only reduced feature variables as input for the evaluation work, this study also compares the predictive effects of the model when all feature variables and only reduced feature variables are used. The GBDT prediction model proposed in this paper predicted melon yield, sugar content, and hardness using reduced features as input, and the model R² could reach more than 90%. Therefore, this method can effectively help growers carry out early non-destructive inspection and growth prediction of melons in the field.

China has become the world's top producer and consumer of melons. 48.7% of the world's melon production in 2020, and per capita consumption is 2.6 times higher than the world average. Beijing, Tianjin, and Hebei are well-known melon-producing areas in China, where melons are crisp, watery, sweet, and well-loved by people. The temperature difference between morning and evening in the north is significant, which helps melons accumulate sugar. The three places' climate conditions and production facilities have their characteristics, and there is a broad space for collaboration in product supply and marketing, technical cooperation, and other aspects.

While breeding new varieties, traits such as growth, flowering, fruiting, fruit size, and color are related to the selection of superior individuals. They are closely related to yield and quality. During melon growth, quality depends not only on its shape and appearance characteristics but also on its soluble solids content, hardness, total acidity, and ripeness, among other factors. In the melon harvesting process, the fruit's weight and maturity are considered the comprehensive evaluation index in the sales process, and the detection of the maturity can protect consumers' interests and improve the processing quality and production-grade of melon to increase market competitiveness. There is a strong demand in the agricultural field to predict its yield and quality. Correlation analysis and prediction of features affecting melon yield and quality can facilitate early decision-making by farmers or plant breeders or enable producers to predict financing aspects before crop harvest.

Machine learning, a branch of AI (Artificial Intelligence), is a practical approach to providing better yield and quality predictions based on multiple features. It can automatically process the relationship between input and output variables and mine the implicit patterns from example samples to 'learn' the structural description of these data. Machine learning algorithms can automatically solve significant nonlinear problems and support better decision-making and operations in real-world scenarios without human intervention. Gradient Boosted Decision Tree, as a mature integrated learning algorithm, focuses on effectively reducing the deviation of predicted values from the actual ones. It connects multiple regression tree models in series to form a strong learner.
and its base-learner regression tree model has the characteristics of high efficiency and insensitive missing values. The performance is improved by continuously fitting the residuals of the previous tree\textsuperscript{11}, which focuses more on the accuracy of the learning model\textsuperscript{12}. It also has the advantages of high efficiency, accurate prediction, insensitivity to the original data, and high interpretability of the model\textsuperscript{13,14}. Gradient Boosted Decision Tree models have been used to predict stress, soil water content, and remote sensing images in agriculture. Zhang et al. used the GBDT model to build a two-stage identification model for 14 features in high-resolution images of early stress in maize, which can be well applied not only for moisture stress detection in non-destructive measurements but also in the field\textsuperscript{15}. Liu et al. established a GBDT model for Soil Moisture (SM) inversion by satellite, and the study showed the excellent performance of GBDT in Soil Moisture Active Passive (SMAP) SM downscaling\textsuperscript{16}. Wu et al. used a GBDT model for simulating monthly mean daily ETo (evapotranspiration) using only local or cross-site temperature data. They showed higher estimation accuracy than other models in local applications\textsuperscript{17}. Yang et al. applied GEOBIA (Geographic Object-Based Image Analysis) to crop remote sensing. GBDT and RF have higher classification accuracy and less computation time than SVM and thus are considered more suitable for crop classification that requires a large number of image features\textsuperscript{18}.

In this paper, data on fruit phenological features, shape features (including appearance and internal features), and color features (including exocarp and endocarp color features) were recorded during the growth of melons. The GBDT algorithm was used to analyze and predict the importance of features affecting melon fruits’ yield, sugar content, and hardness values. The prediction performance and results of two classical regression models, Support Vector Regression (SVR) and Random Forest (RF), are comprehensively compared. Also, the prediction results of the model when all feature variables and only reduced feature variables are compared so that the model can be better applied to non-destructive testing tasks in the field.

Materials and methods
Culture implementation and maintenance. The study was conducted at the experimental base of Tianjin Academy of Agricultural Sciences. It was located in Beichen District, Tianjin City, on the east coast of Asia and Europe at mid-latitude. N: 39°18′11″ E: 117°13′23″. The experiment was carried out in the summer and autumn in the Tianjin area, with temperatures ranging from 16 to 25 °C from August to October, average precipitation of 214 mm, and 59% sunshine percentage in the area. Seedlings were planted on August 5 and harvested on October 16.

The test crop was the thick-skinned melon variety “232”, provided by the Institute of Vegetables, Tianjin Academy of Agricultural Sciences. The variety ripened about 35 d after flowering and about 15 d after the melon expansion period. Melon seedlings were transplanted when they had two true leaves and one terminal bud at a planting density of 33,945 plants/ha. Organic fertilizer and high potassium compound fertilizer were applied sufficiently before planting, mixing, and rototilling.

We confirm that the plants used in the article were sampled and measured during field production and comply with international, national and institutional guidelines for this study.

Measurement methods and instruments. All indicators were evaluated at harvest. Electronic scales were used to weigh weight indicators, vernier calipers to measure shape indicators, hardness meters to measure flesh hardness indicators, sugar meters to measure sugar content, and color indicators were measured using a colorimeter for indicator measurement, and phenological information was recorded manually.

The measuring instruments and specifications used in this study mainly include Small-range precision balance (ZG-TP203), specification 5 kg/0.01 g, Shanghai Ranhao Electronics Co., Ltd; Vernier caliper (range: 0 ~ 150 mm), Shenzhen San En Shi (3 NH) Technology Co. GY-4 digital display fruit hardness tester, Beijing Jinkolida Electronic Technology Co., Ltd; PAL-1 fruit digital display sugar tester, Japan ATAGO automatic portable colorimeter, Shenzhen San En Shi (3 NH) Technology Co. GY-4 digital display fruit hardness tester, Beijing Jinkolida Electronic Technology Co., Ltd; PAL-1 fruit digital display sugar tester, Japan ATAGO Co. as shown in Fig. 1.

Each complete sample was collected three times for weight index to take the average value. Sugar content value was obtained at the edge of the seed cavity and endocarp, and each half of the sample was collected three times to take the tie-breaker value. Hardness values were collected at the center of the endocarp, and each half of the sample was collected simultaneously on the left and right sides to take the average value. The melon fruit develops from the receptacle and the ovary. The fruit can be divided into the pericarp and the seed cavity. The pericarp is composed of the exocarp, mesocarp, and endocarp, and the mesocarp and endocarp are not clearly defined. Both consist of large, thin-walled cells rich in water and soluble sugars. The shape indexes were measured by vernier calipers for the corresponding position indexes. The description of each index is shown in Table 1.

Gradient boosting decision tree model. Gradient Boosted Decision Tree, an integrated learning algorithm with CART Classification and regression tree (CART) as the base model, is composed of two parts: Decision Tree and Gradient Boosting. Its primary learner is the regression tree, which minimizes the mean square error by constructing a function to fit the elements in the data set. Multiple regression tree models are composed using a combination strategy to obtain the GBDT integrated learner.

A GBDT learning model was developed by combining phenological features, shape features, and color features with melon fruit yield, sugar content, and hardness.

where, $x_i, y_i$ are the corresponding criterion parameters; $S_i$ is the corresponding leaf node in the regression tree model. $D = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \}$ is a dataset containing $m$ training samples, each described by $d$ feature attributes, i.e., $x_i = [x_{i1}, x_{i2}, \ldots, x_{id}]$; $y_i$ is the yield, sugar content, and hardness of the corresponding sample. The basic algorithmic flow of the established GBDT model is as follows\textsuperscript{19}.
(1) Initialize the base learner \( f_0(x) \):

\[
f_0(x) = \arg \min_a \sum_{j=1}^m L(y_j, a)
\]

In which: \( L(y_j, a) \)—loss function; \( \arg \min \)—the function that determines the value of \( a \) taken at the minimum of the value of the loss function.

---

**Table 1.** Feature variable related description.

| Number | Feature name                  | Measurement meaning                                                                 | Abbreviation |
|--------|-------------------------------|-------------------------------------------------------------------------------------|--------------|
| Phenological features          |                               |                                                                                     |              |
| 1      | Days of flowering             | Affects fruit ripening and time to market                                            | DF           |
| 2      | Number of days to fruit ripening |                                                                              | NF           |
| Appearance features            |                               |                                                                                     |              |
| 3      | Fruit width                   | Influence the appearance of commercial-ity and fruit weight\(^{19} \)                   | FW           |
| 4      | Fruit length                  |                                                                                     | FL           |
| 5      | Fruit shape index             | Impact on the appearance of commodity classification\(^{20} \)                   | FSI          |
| Internal features               |                               |                                                                                     |              |
| 6      | Sum of exocarp and mesocarp fruit widths | Inedible parts                                                                 | SWI          |
| 7      | Endocarp thickness            | Edible parts                                                                         | ET           |
| 8      | Total peel thickness          | Total flesh thickness, excluding the seed cavity                                       | TPT          |
| 9      | Seed cavity width             | Size of seed cavity affects weight per fruit; edible portion                         | SCW          |
| 10     | Seed cavity length            |                                                                                     | SCL          |
| Exocarp color features         |                               |                                                                                     |              |
| 11 – 15 | \( L^a b^c h^* \) color space of exocarp | Influence on the determination of fruit ripeness\(^{21} \)                       | XC\(_L^a\), XC\(_a^b\), XC\(_b^c\), XC\(_h^*\) |
| 16 – 18 | Hunter \( L^a b^c h^* \) color space of exocarp |                                                                                     | XC\(_HL^a\), XC\(_Ha^b\), XC\(_Hb^*\) |
| 19 – 21 | \( XYZ \) color space of exocarp |                                                                                     | XC\(_X\), XC\(_Y\), XC\(_Z\) |
| Endocarp color features        |                               |                                                                                     |              |
| 22 – 26 | \( L^a b^c h^* \) color space of Endocarp | Influence on the determination of fruit ripeness                               | NC\(_L^a\), NC\(_a^b\), NC\(_b^c\), NC\(_h^*\) |
| 27 – 29 | Hunter \( L^a b^c h^* \) color space of Endocarp |                                                                                     | NC\(_HL^a\), NC\(_Ha^b\), NC\(_Hb^*\) |
| 30 – 32 | \( XYZ \) color space of Endocarp |                                                                                     | NC\(_X\), NC\(_Y\), NC\(_Z\) |

---

**Figure 1.** Melon fruit shape features.
(2) A series of CART regression trees are built, on which the residuals are fitted using a gradient boosting technique. In the k th (k = 1,2,...,K) iteration, for each sample (xj,yj), GBDT specifies the negative gradient of the loss value as the residual estimate. In this paper, we choose Least mean squared error, as the loss function.

\[ L(y_j,f(x_j)) = (y_j - f(x_j))^2 \]  

(2)

The residual estimates are:

\[ R_{jk} = -\frac{\partial L(y_j,f(x_j))}{\partial f(x_j)} = y_j - f(x_j) \]  

(3)

(3) After determining the residual estimates, the CART regression tree is fitted to obtain the leaf node region of the k th tree as \( c_jk \) (j = 1, 2, ... , f), and f is the number of leaf nodes in the regression tree. For each leaf node region, determine the best-fit value \( \beta_{jk} \) that minimizes the corresponding loss function.

\[ \beta_{jk} = \arg \min_\beta \sum_{x_j \in c_{jk}} L(y_j,f_k-1(x_j) + \beta) \]  

(4)

\[ \beta_{jk} = \arg \min_\beta \sum_{x_j \in c_{jk}} (y_j - f_k-1(x_j) - \beta)^2 \]  

(5)

(4) Update Learner \( f_k(x) \):

\[ f_k(x) = f_{k-1}(x) + \sum_{j=1}^{f} \beta_{jk} \cdot \eta \]  

(6) in which: \( \eta \)—learning rate.

(5) After the end of the iteration, a GBDT strong learner is formed, which can be expressed as:

\[ F(x) = f_0(x) + \sum_{k=1}^{K} \sum_{j=1}^{f} \beta_{jk} \cdot \eta \]  

(7)

(6) The principle of feature importance calculation can be expressed in the following way13: in the description of the approximate \( F(x) \), the relative impact of individual input \( x_j \) on the change in \( F(x) \) over the distribution of joint input variables \( I_j \) can be expressed as:

\[ I_j = \left( \text{Ex} \left[ \frac{\partial F(x)}{\partial x_j} \right]^2 \cdot \text{var}_{x_j} \right)^{1/2} \]  

(8)

The global importance of a feature is measured by the average of the importance of the feature in a single tree:

\[ \hat{I}_j(T) = \frac{1}{M} \sum_{m=1}^{M} \hat{I}_m^2(T_m) \]  

(9)

In which: \( M \)—the number of trees.

The importance of features in a single tree is as follows:

\[ \hat{I}_j(T) = \sum_{T_m}^{L-1} \hat{I}_m^2(T_{m=1}) \cdot 1(\nu_t = j) \]  

(10)

In which: \( L \)—the number of leaf nodes of the tree, \( L - 1 \) is the number of non-leaf nodes of the tree (the constructed trees are binary trees with left and right leaves); \( \nu_t \)—the feature associated with node t; \( \hat{I}_m^2(T) \)—the reduced value of the squared loss after node splitting.

**Experimental results and analysis**

For GBDT, this paper uses GS (grid search method) to optimize the four hyperparameters on the modeled dataset. Maximum number of iterations is 750, maximum depth is 5, learning rate is 0.05, minimum number of leaves is 1.

**Evaluation metrics.** The prediction results of each model were evaluated and compared using the \( R^2 \) (R-squared, goodness-of-fit). Meanwhile, the MAE (Mean absolute error) and RMSE (root mean squared error) metrics were used to comprehensively compare the prediction effects of each model. The TIME (Consumption time) of the model calculation is used to evaluate the efficiency of the model calculation. Among them, MAE reflects the deviation between predicted and actual values, and RMSE reflects the standard deviation of the difference between predicted and actual values.
In which:

- $y$ — the average of the target;
- $y_j$ — the fruit weight, sugar content and hardness values as noted in the field;
- $p_j$ — the predicted value of the $j$th sample estimated by the model.

Importance analysis of feature variables. The importance of each feature variable in the prediction process of melon yield, sugar content, and hardness value was calculated according to the principle of feature importance, as shown in Fig. 2.

Importance analysis was performed according to the importance of the three target feature variables and categories. For yield: Fruit width (FW), Fruit length (FL), Seed cavity length (SCL), Total peel thickness (TPT), and internal character parameters related to endocarp and seed cavity were decisive for yield. For sugar content, the parameters Endocarp thickness (ET), Seed cavity width (SCW), Sum of exocarp and mesocarp fruit widths (SWI), and Total peel thickness (TPT), which are related to the size of the edible part of the fruit, are closely related. For hardness, the internal characteristics Total peel thickness (TPT), Endocarp thickness (ET), Sum of exocarp and mesocarp fruit widths (SWI), and exocarp color features have a significant influence on the hardness index. Phenological characteristics have a more critical effect on sugar content than yield and hardness. Among the color indicators, $XC_{Hb^*}$ and $XC_{b^*}$, which were spatially related to the yellow-blue color characteristic of the exocarp, had significant associations for both sugar content and hardness, and the exocarp color was more important for all three indicators compared to the endocarp color.

According to Fig. 3. The essential characteristics for yield, sugar content, and hardness were Fruit width (FW), Endocarp thickness (ET), and Total peel thickness (TPT). Fruit width (FW) had more influence on yield and sugar content than other features, and ET affected both sugar content and hardness. The influencing factors of yield index were mainly concentrated on appearance features, and the influencing factors of sugar content and hardness index were mainly concentrated on internal features. At the same time, the color features of outer peel also had a relatively significant influence on them.

\[
R^2 = 1 - \frac{\sum_{j=1}^{T} (y_j - p_j)^2}{\sum_{j=1}^{T} (y_j - \bar{y})^2} \tag{11}
\]

\[
MAE = \frac{1}{T} \sum_{j=1}^{T} |y_j - p_j| \tag{12}
\]

\[
RMSE = \sqrt{\frac{1}{T} \sum_{j=1}^{T} (y_j - p_j)^2} \tag{13}
\]

In which: $\bar{y}$ — the average of the target; $y_j$ — the fruit weight, sugar content and hardness values as noted in the field; $p_j$ — the predicted value of the $j$th sample estimated by the model.

**Importance analysis of feature variables.**

![Figure 2. Importance order of feature variables related to yield, sugar content, and hardness.](image-url)
Predictive effect analysis. The GBDT learning model developed in this paper predicts melon yield, sugar content, and hardness values for 32 feature variables, including phenological features, appearance features, internal features, exocarp color features, and endocarp color features, respectively. With the 255 samples obtained, the learning model randomly divides the training and testing sets in a 7:3 ratio by setting a random seed number in the training and testing phases. The evaluation metrics are shown in Table 2. The prediction results are shown in Fig. 4.

Discussion
In order to facilitate growers to be able to operate in the field and help them to estimate them in the early stages of planting, a reduced feature prediction model with melon phenological features, appearance features, and exocarp color features as inputs were constructed and discussed. The purpose is used to compare the feasibility of using prediction models for melon yield, sugar content, and hardness under non-destructive conditions.

At the same time, to evaluate and test the prediction effects of GBDT learning models on melon yield, sugar content, and hardness, this paper establishes GBDT, SVR, and RF simultaneously and with constant input and output feature parameters and sample data set division. The comprehensive analysis to compare the prediction results is shown in Fig. 5, and the evaluation indexes are shown in Table 3. R2, RMSE, MAE, and TIME were used to predict the results of each model. The RF parameters: the number of iterations is 750, the maximum depth is 5, and the minimum number of leaf nodes is 1. The SVM parameters: the kernel uses ‘RBF’, C is 100, and the learning rate is 0.001.

The comparison shows that since the growth process of decision trees is to select and partition features continuously, GBDT, which consists of a large number of decision trees, has the inherent advantage that the importance ranking of features can be easily obtained and is highly interpretable. The Support Vector Regression (SVR) model is used for regression using the idea of Support Vector Machine (SVM), which allows for an ε distance between the predicted and true values of the model to improve generalization compared to traditional regression models, and has an advantage for nonlinear24. Gradient Boosted Decision Tree and Random Forest (RF) final results are determined by multiple trees together, and different from RF is the idea of training the base learner. RF uses the Bagging method, which cannot improve the bias. In contrast, GBDT uses Boosting method, and each iteration is weighted to the sample based on the prediction result of the previous iteration. As iterations continue, low bias can be guaranteed to improve the model’s generalization ability.

As shown in Fig. 6, by comparing the R2 indicators of the three models with all the feature variables and the reduced feature variable as inputs, it can be seen that the GBDT model has a higher R2 for yield prediction than the sugar content and hardness models before and after the feature reduction. In predicting yield using the three different models, the use of reduced features improved the R2 of the model compared to all feature variables as input. On the contrary, reducing features as input decreased the R2 of all three prediction models in predicting sugar content and hardness, and the SVR model predicted a significant decrease in R2 of more than 30%. It related to the importance of all the feature variables in the prediction models discussed earlier, where reduced features were mostly located at the top of the variable importance ranking in the yield prediction models, and FW and

Table 2. Predictive metrics.

|                | RMSE  | MAE   | R²   | TIME  |
|----------------|-------|-------|------|-------|
| Yield forecast | 0.041421 | 0.001716 | 0.999048 | 1.0118  |
| Sugar content prediction | 0.274511 | 0.075356 | 0.976626 | 0.909321  |
| Hardness prediction | 0.014827 | 0.121767 | 0.993375 | 0.853631  |
FL were close to 100% in the feature importance scores. In contrast, more internal features, such as ET, TPT, and SCW, in the sugar content and hardness prediction models had higher importance scores. Adding some internal fruit measures could further improve the prediction accuracy of sugar content and hardness.

The GBDT prediction model proposed in this paper was used to predict melon yield, sugar content, and hardness using reduced features as input, and the model R2 could reach more than 90%. Therefore, using non-destructive feature datasets to predict melon fruit yield, sugar content, and hardness can achieve high prediction accuracy, and the method is feasible.

Figure 4. GBDT prediction effectiveness.
Conclusions

In this paper, we propose a machine learning method using GBDT and GS (Grid Search Method) optimization-seeking hyperparameter method for melon varieties in northern China to analyze the importance of the characteristic variables affecting melon yield, sugar content, and hardness and to successfully predict the test data. A significant benefit of early forecasting is that it allows early assessment and preparation for those involved in breeding new varieties, making production decisions, and seeking financial support. To improve the model's usefulness, combined with hardware devices for use in the field without destroying the fruit itself, this study

Figure 5. Comparison of yield (a), sugar content (b), and hardness (c) prediction results of different models with the reduced features as inputs.
screened non-destructive features of melon fruit, including melon phenological features, appearance features, and exocarp color features, to construct a non-destructive prediction model. Through the analysis of the study, the model still has acceptable prediction results compared to all features used as input. At the same time, the model training time decreases, and the running efficiency improves due to the reduction of feature variables. Therefore, in the future, by embedding the prediction model into the vision inspection equipment, it will be easy to achieve non-destructive, accurate, and fast prediction and evaluation work in the field.

Data availability
The availability of some or all of the data generated or analyzed in this study is restricted and the entire data cannot be made public until the variety is officially licensed as a variety. It is not publicly available, but are available from the corresponding author on reasonable request.

Received: 17 July 2022; Accepted: 19 September 2022
Published online: 28 October 2022

References
1. Sun, L. X., Wang, X. J., Wu, J. X. & Mao, S. P. Research on regional distribution changes and driving factors of watermelon and melon production in China. Chin. J. Agric. Resour. Region. Plan. https://kns.cnki.net/kcms/detail/11.3513.S.20220705.1954.008.html (2022).
2. Wang, K. High efficient cultivation techniques of muskmelon in winter and spring in solar. J. Agric. Catastrophol. 11(3), 1 (2021).
3. Schwarzacher, H. H. Domestication, genomics, and the future for banana. Ann. Bot. 100(5), 1073–1084 (2007).
4. Sun, J. T. Study on Nondestructive Dection of Hami Melon Maturity Based on Information Fusion of Spectrum and Image (Shihezi University, 2017).
5. Lv, C. et al. Research progress of melon ripeness detection technology. Jiangsu Agric. Sci. 42(1), 244–246 (2014).
6. Zhang, X., Chen, Y. & Xiao, G. A review of crop yield prediction based on machine learning. Anhui Agri. Sci. Bull. 27(03), 1 (2021).
7. Soares, J., Pasquall, M., Lacerda, W. S., Silva, S. O. & Donato, S. Utilization of artificial neural networks in the prediction of the bunches’ weight in banana plants. Sci. Hortic. 155, 24–29 (2013).
8. Klompenburg, T. V., Kassahun, A. & Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* 177, 105709 (2020).
9. McQueen, R. J., Garner, S. R., Nevill-Manning, C. G. & Witten, I. H. Applying machine learning to agricultural data. *Comput. Electron. Agric.* 12(4), 275–293 (1995).
10. Chlingaryan, A., Sukkarieh, S. & Whelan, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Comput. Electron. Agric.* 151, 61–69 (2018).
11. Pang, S., Yang, X. Y., Zhang, X. F. & Lin, X. S. Fault diagnosis of rotating machinery with ensemble kernel extreme learning machine based on fused multi-domain features. *ISA Trans.* 98, 320–337 (2020).
12. Zhou, Z. H. H. *Machine Learning* 100–123 (Tsinghua University Press, 2016).
13. Friedman, J. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29, 203451 (2001).
14. Su, X. H., Sun, J. M., Gao, X. & Wang, M. Prediction method of drilling rate of penetration based on GBDT algorithm. *Comput. Appl. Softw.* 36(12), 87–92 (2019).
15. Zhuang, S., Wang, P., Jiang, B., Li, M. & Gong, Z. Early detection of water stress in maize based on digital images. *Comput. Electron. Agric.* 140, 461–468 (2017).
16. Liu, Y. et al. Downsampling satellite retrieved soil moisture using regression tree-based machine learning algorithms over Southwest France. *Earth Space Sci.* 7, e2020001267 (2020).
17. Wu, L., Peng, Y., Fan, J. & Wang, Y. Machine learning models for the estimation of monthly mean daily reference evapotranspiration based on cross-station and synthetic data. *Hydrol. Res.* 50(6), 1730–1750 (2019).
18. Yang, L., Mansaray, L., Huang, J. & Wang, L. Optimal segmentation scale parameter, feature subset and classification algorithm for geographic object-based crop recognition using multisource satellite imagery. *Remote Sens.* 11(5), 514 (2019).
19. Naroufi Rad, M. R., Ghazandeh, A. & Abbas Koohpaygani, J. Predicting eggplant individual fruit weight using an artificial neural network. *Int. J. Veg. Sci.* 2017, 1290001 (2017).
20. He, J., Ma, Z. H. Y., Chu, X., Liu, H. L. & Xiao, T. Y. Research on mango shape evaluation method based on machine vision. *Mod. Agric. Equip.* 42(01), 56–60 (2021).
21. Xu, D. F., Zhao, H. M., Xu, J. D. & Zhang, S. J. Study on the maturity discrimination of muskmelon based on machine vision. *Farm Prod. Process.* 2020(14), 19-22,25 (2020).
22. Wang, J. M., Ye, Y. R., Rao, C. H. M. & Zhao, R. J. Prediction on the composite shear strength of ceramics lightweight aggregate and normal concrete based on GBDT algorithm. *J. Build. Mater. Tech.* 2022, 1–10 (2022).
23. Cui, H., Yu, X. L., Pang, J. W. & Yang, S. H. S. H. Influential BOD soft sensing models based on BP-ANN and improved SVR. *J. Harbin Inst. Technol.* 54(02), 59–66 (2022).
24. Zhang, H. Y., Zhang, Y., Li, M. Z. & Li, X. H. BSO-SVR-based remote sensing time-series yield estimation model for banana. *Trans. Chin. Soc. Agric.* 52(S1), 98–107 (2021).

**Acknowledgements**
This work was supported by Key Technology Research and Development Program of Hebei Province [grant numbers: 22327213D]; The Key Laboratory of Storage of Agricultural Products, Ministry of Agriculture and Rural Affairs [grant numbers: Kf2021003]; Natural Science Foundation of Hebei Province [grant numbers: 202102136]; Tianjin Science and Technology Plan Project [grant numbers: 21ZYCGSN00320].

**Author contributions**
C.Q.: Conceptualization, Methodology, Software, Validation, Writing—Original Draft. T.D.: Writing - Review & Editing, Supervision. S.S.: Validation, Investigation. J.W.: Supervision. W.L.: Data curation, Validation. H.Z.: Investigation, Resources.

**Competing interests**
The authors declare no competing interests.

**Additional information**
Correspondence and requests for materials should be addressed to C.Q.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s) 2022