sustainability

MDPI

Article

A Crop Harvest Time Prediction Model for Better Sustainability, Integrating Feature Selection and Artificial Intelligence Methods

Shu-Chu Liu [1,*], Quan-Ying Jian [1], Hsien-Yin Wen [1] and Chih-Hung Chung [2]

[1] Department of Management Information Systems, National Pingtung University of Science and Technology, Pingtung 912301, Taiwan
[2] Department of Educational Technology, Tamkang University, New Taipei City 251301, Taiwan
* Correspondence: sliu@mail.npust.edu.tw

Abstract: Making an accurate crop harvest time prediction is a challenge for agricultural management. Previous studies of crop harvest time prediction were mainly based on statistical methods, and the features (variables) affecting it were determined by experience, resulting in its inaccuracy. To overcome these drawbacks, the objective of this paper is to develop a novel crop harvest time prediction model integrating feature selection and artificial intelligence (long short-term memory) methods based on real production and climate-related data in order to accurately predict harvest time and reduce resource waste for better sustainability. The model integrates a hybrid search for feature selection to identify features (variables) that can effectively represent input features (variables) first. Then, a long short-term memory model taking the selected features (variables) as input is used for harvest time prediction. A practical case (a large fruit and vegetable cooperative) is used to validate the proposed method. The results show that the proposed method (root mean square error (RMSE) = 0.199, mean absolute percentage error (MAPE) = 4.84%) is better than long short-term memory (RMSE = 0.565; MAPE = 15.92%) and recurrent neural networks (RMSE = 1.327; MAPE = 28.89%). Moreover, the nearer the harvest time, the better the prediction accuracy. The RMSE values for the prediction times of one week to harvesting period, two weeks to harvesting period, three weeks to harvesting period, and four weeks to harvesting period are 0.165, 0.185, 0.205, and 0.222, respectively. Compared with other existing studies, the proposed crop harvest time prediction model, LSTMFS, proves to be an effective method.

Keywords: a crop harvest time prediction model; feature selection; artificial intelligence; long short-term memory; sustainability

Citation: Liu, S.-C.; Jian, Q.-Y.; Wen, H.-Y.; Chung, C.-H. A Crop Harvest Time Prediction Model for Better Sustainability, Integrating Feature Selection and Artificial Intelligence Methods. *Sustainability* **2022**, *14*, 14101. https://doi.org/10.3390/su142114101

Academic Editor: Teodor Rusu

Received: 30 September 2022
Accepted: 26 October 2022
Published: 28 October 2022

1. Introduction

Making an accurate crop harvest time prediction is a challenge for sustainable agricultural management, but it could eventually decrease resource waste [1]. For harvest time prediction, previous studies have attempted to use statistical analyses to make predictions [2–4]. In recent years an increasing number of scholars have been utilizing artificial intelligence (AI) to solve the problem of harvest time prediction [1,5,6], and their results show that artificial intelligence methods are better than statistical methods [1]. For example, the long short-term memory (LSTM) model is used to construct a relationship between data by integrating data from different sources for machine learning (ML) [7,8]. In addition, essential features (variables) are mainly determined empirically, but they are not based on crop characteristics and real data, resulting in relatively poor prediction accuracy [9]. To overcome these drawbacks, the objective of this paper is to develop a novel crop harvest time prediction model integrating feature selection and artificial intelligence methods based on real production and climate-related data in order to accurately predict harvest time and reduce resource waste for better sustainability.

2. Contribution

Crop harvest time prediction is an important operation for agricultural management. Previous studies of crop harvest time prediction were mainly based on statistical methods, and the features (variables) affecting it were determined by experience, resulting in its inaccuracy. To overcome these drawbacks, this paper develops a novel and effective crop harvest time prediction model, LSTMFS, which integrates a hybrid search for feature selection and a long short-term memory (LSTM) model based on real production and climate-related data, in order to accurately predict harvest time and reduce resource waste for better sustainability.

3. Literature Review

3.1. Artificial Intelligence for the Crop Harvest Time Prediction Model

Many studies have attempted to predict crop harvest time using statistical analyses or time series analyses in the past [10], and some scholars have recently used artificial intelligence (AI) methods to solve the problem of crop harvest time prediction [5]. de Souza et al. [6] proposed artificial neural networks (ANNs) for predicting banana harvest time. Furthermore, recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been adopted to predict crop harvest time [1,11]. However, ANNs and CNNs (more suitable for image data) cannot handle the pre–post relationship with the presentation of time series. For time series, scholars have used RNNs or LSTM to control the pre–post relationship between data, as their performance is better than that of ANNs and CNNs [12]. According to previous studies, the LSTM prediction model is a modified version of the RNN and has better accuracy than that of the RNN [13,14]; therefore, the following discussion focuses on the LSTM prediction model. For example, Zhang et al. [12] used sensors to collect various real-time information and trained LSTM to predict the machine's durability. Sagheer and Kotb [15] proposed a deep LSTM network for time series to predict oil production. Karevan and Suykens [7] adopted LSTM to forecast weather. Yadav et al. [16] adopted LSTM to forecast the Indian stock market. Chimmula and Zhang [17] adopted LSTM to predict the propagation trend of COVID-19 in Canada and received alerts before the crisis occurred. According to the review above, the LSTM model is used as a prediction model because crop harvest time is a type of time series, which is also suitable for LSTM to predict.

3.2. Feature Selection Method

The selection of features (variables) for the crop harvest time prediction model has not been explored in the past. Most features are decided by empirical rules, which reduces the model's accuracy (features selected empirically cannot change with real production and climate-related data, resulting in relatively poor prediction accuracy [9]). The selection of features (variables) is based on the selection of some features (variables) that can effectively represent input features (variables) and reduce the effect of uncorrelated variables, thereby allowing better prediction results to be obtained [18]. Feature (variable) selection is mainly conducted via one of two methods: (1) the filter method and (2) the wrapper method. The filter method determines each variable's score based on the variable's importance and sets a threshold value. However, this method ignores the mutual influence of variables, and this way of deciding on the variables makes the prediction results relatively poor. The wrapper method considers all variables in the wrapper at the same time to overcome the problem of the filtering method. However, the best solution lies in a combination of problems (2^N). The search method can be divided into two main types: the regional search and the evolutionary search. The regional search method is mainly based on regional solutions (such as variable neighborhood search (VNS), taboo search (TS), and simulated annealing (SA)), which can search different dimensions quickly and express problems easily [19]. The most commonly adopted methods include VNS [20], TS [21], and SA [22]. Related studies have shown that VNS is superior to TS and SA [23]. In recent years, particle swarm algorithms (PSO) (evolutionary search methods) have been frequently mentioned

in feature selection problems, and they are superior to region-based methods and genetic algorithms (GA) [24] but have the drawback of converging too quickly. Hybrid algorithms are increasingly being used to solve feature selection problems and have been proven to be superior to single algorithms. Moradi and Gholampour [25] proposed a hybrid method for feature selection, which integrates regional searches with PSO for full domain searches and outperforms the GA, SA, ACO, and PSO methods. Esfandiarpour-Boroujeni et al. [26] used a hybrid particle swarm optimization–imperialist competitive algorithm-supported vector regression method to predict apricot production and identified 18 out of 61 variables as the best features (variables) for predicting apricot production. Li and Becker [27] proposed a mixed-mode integration of LSTM with the feature selection method of PSO for predicting tariffs. Based on the discussion above, this paper proposes a hybrid search. The search uses the PSO to search for the best solution. To avoid fast convergence, PSO integrates VNS to skip the best solution in the region.

4. Materials and Methods

4.1. Crop Harvest Time Prediction Model

In recent years, there have been no comprehensive studies on the distinguishing features affecting maturity (based on Elsevier Science, Springer-Verlag, EBSCO, ProQuest, and Google search results), but there has been research on individual features affecting growth; for instance, (1) Hatfield and Prueger [28] showed that temperature and accumulated temperature significantly affect plant growth; (2) Punia et al. [29] showed that solar radiation affects plant growth; (3) Ndamani and Watanabe [30] showed the effect of rainfall on plant growth; (4) Hirai et al. [31] showed the effect of humidity on plant growth; (5) Gardiner et al. [32] illustrated the relationship between wind speed and plant growth. According to the aforementioned literature, the considered features include cumulative accumulated temperature (according to Ref. [1]), accumulated temperature (according to Ref. [1]), accumulated sunshine hours (according to Ref. [2]), accumulated total sky radiation (according to Ref. [2]), accumulated radiation (according to Ref. [2]), cumulative rainfall (according to Ref. [3]), cumulative precipitation hours (according to Ref. [3]), average humidity (according to Ref. [4]), and average wind speed (according to Ref. [5]). The influential features associated with the crop harvest time prediction model were compiled as follows: (1) cumulative accumulated temperature, (2) accumulated temperature, (3) accumulated sunshine hours, (4) accumulated total sky radiation, (5) accumulated radiation, (6) accumulated rainfall, (7) cumulative precipitation hours, (8) average humidity, and (9) average wind speed.

The data from three days of lag for the 9 selected features mentioned above are considered as input (After testing for n days of lag, the input variables use the values of previous observations ($t - n, \ldots., t - 2, t - 1$) at time t. These previous observations are called lags one, two, and three. The RMSE (=0.565) of three days of lag for LSTM is the smallest, and it is selected for harvest time prediction.). A total of 27 variables are used as input x_t at time t for LSTM, and the output variable y_t at time t is the harvest time (the number of days until harvesting from time t). Details of the variables are compiled in Table 1.

Table 1. Variables used for long short-term memory (LSTM).

Input Variables x_t	**Output Variable y_t**
(1) Cumulative accumulated temperature one day ago	The number of days until harvesting from time *t*
(2) Cumulative accumulated temperature two days ago	
(3) Cumulative accumulated temperature three days ago	
(4) Accumulated temperature one day ago	
(5) Accumulated temperature two days ago	
(6) Accumulated temperature three days ago	
(7) Accumulative sunshine hours one day ago	
(8) Accumulative sunshine hours two days ago	
(9) Accumulative sunshine hours three days ago	
(10) Accumulated total sky radiation one day ago	
(11) Accumulated total sky radiation two days ago	
(12) Accumulated total sky radiation three days ago	
(13) Accumulated radiation one day ago	
(14) Accumulated radiation two days ago	
(15) Accumulated radiation three days ago	
(16) Accumulated rainfall one day ago	
(17) Accumulated rainfall two days ago	
(18) Accumulated rainfall three days ago	
(19) Cumulative precipitation hours one day ago	
(20) Cumulative precipitation hours two days ago	
(21) Cumulative precipitation hours three days ago	
(22) Average humidity one day ago	
(23) Average humidity two days ago	
(24) Average humidity three days ago	
(25) Average wind speed one day ago	
(26) Average wind speed two days ago	
(27) Average wind speed three days ago	

The data from the previous three days for the 9 selected features mentioned above are considered (after testing, three days is the best parameter). A total of 27 variables (features) are used as input x_t at time t for long short-term memory (LSTM) (i.e., cumulative accumulated temperature one day ago, cumulative accumulated temperature two days ago, cumulative accumulated temperature three days ago, accumulated temperature one day ago, accumulated temperature two days ago, accumulated temperature three days ago, accumulative sunshine hours one day ago, accumulative sunshine hours two days ago, accumulative sunshine hours three days ago, accumulated total sky radiation one day ago, accumulated total sky radiation two days ago, accumulated total sky radiation three days ago, accumulated radiation one day ago, accumulated radiation two days ago, accumulated radiation three days ago, accumulated rainfall one day ago, accumulated rainfall two days ago, accumulated rainfall three days ago, cumulative precipitation hours one day ago, cumulative precipitation hours two days ago, cumulative precipitation hours three days ago, average humidity one day ago, average humidity two days ago, average humidity three days ago, average wind speed one day ago, average wind speed two days ago, and average wind speed three days ago), and the output variable y_t at time t is the harvest time (the number of days until harvesting from time t).

The structure of LSTM is shown in Equations (1)–(6). There are four components in the LSTM: a forget gate (f_t), an input gate (i_t), an output gate (o_t), and a memory cell ($\widetilde{C}_t$). This

cell retains values over time intervals, and the three gates are responsible for controlling the flow of information into and out of the cell. At time t, the cell is fed with input x_t and the hidden state h_{t-1} at time $t-1$. The forget gate f_t, the input gate i_t, the output gate o_t, and the memory cell $\widetilde{C}_t$ are calculated as follows:

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \quad (2)$$

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \quad (3)$$

$$\widetilde{C}_t = \tan h(W_c \times [h_{t-1}, x_t] + b_c) \quad (4)$$

where σ and tanh are the sigmoid and hyperbolic tangent activation functions, respectively. The weights and biases of the input gate, output gate, forget gate, and memory cell are denoted by W_i, W_o, W_f, and W_c and b_i, b_o, b_f, and b_c, respectively.

Then, the output cell state C_t and the hidden state h_t at time t can be calculated as follows:

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}_t \quad (5)$$

$$h_t = o_t \times \tan h(C_t) \quad (6)$$

There are five layers for the harvest time prediction in the Keras sequential model (Figure 1): Input layer, one LSTM layer, Dropout layer, Dense layer, and Output layer. In the implementation of the model, the input data x_t in the Input layer include the 27 variables mentioned above. The LSTM layer is adopted with 30 hidden nodes. The activation function used in this layer is a rectified linear unit. A dropout mechanism in the Dropout layer is applied to the inputs of the Dense layer to prevent over-fitting, and the dropout rate is set to 0.4. The Dense layer with a linear activation function is used to return a single continuous value. The adaptive moment estimation function is used in the optimizer parameter. This function defines how the weights of the neural network are updated. The output data y_t in the Output layer represent the predicted harvest time (the number of days until harvesting from time t).

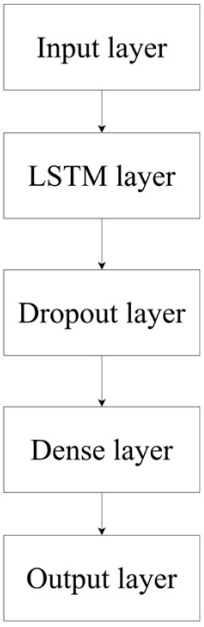

Figure 1. The LSTM model for harvest time prediction.

4.2. Feature Selection Method for the Crop Harvest Time Prediction Model

After the crop harvest time prediction model is determined, the feature (variable) selection method is used to remove some irrelevant input variables in order to improve the accuracy of the prediction. Since the search for the best solution for all variable combinations is a combinatorial problem (complexity is 2^N, where N is the number of all input variables), the proposed method is a hybrid search method integrating a particle swarm search and a large neighborhood search (LNS, a variant of variable neighborhood search (VNS)). First, the parameters are set. Then, the particle position and velocity at the first iteration (generation) are generated (there are psize particles). The particle adaptation values for psize particles are calculated. LNS is executed for the new iteration (generation), and, then, Pbest and Gbest are updated. Has iter reached the default value (=isize)? If it has, (1) stop; otherwise, (2) update the particle velocity and position, (3) execute LNS for Gbest, (4) calculate iter = iter + 1, and (5) go back to calculate the particle adaptation values (Figure 2).

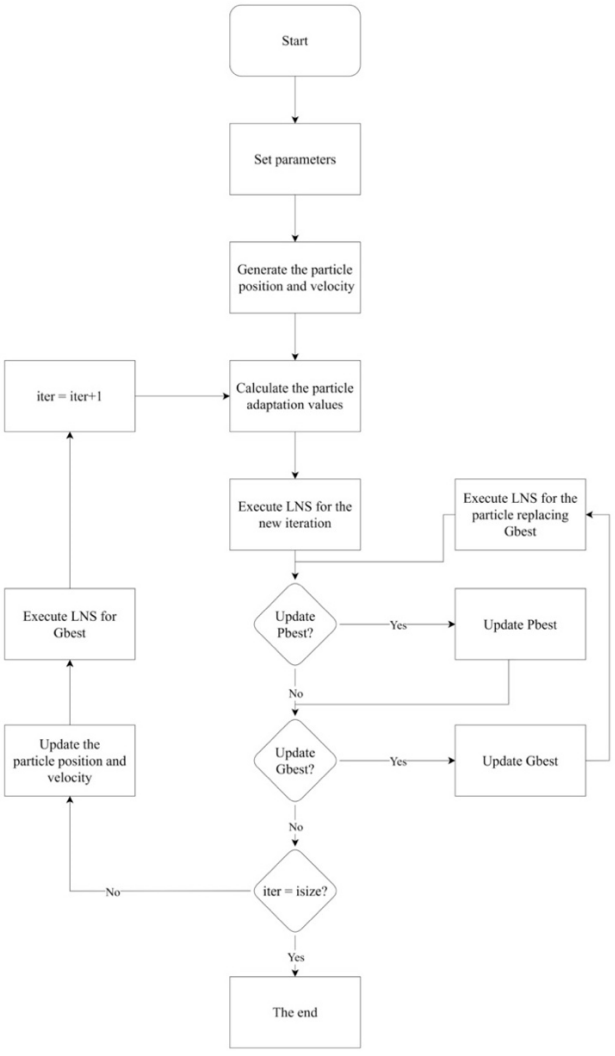

Figure 2. A flowchart of the feature selection method.

4.2.1. Set the Parameters

Set the initial parameters: current iteration (generation) count pointer, iter (=1); current particle pointer, pindex; upper iteration (generation) count, isize; number of particles, psize; inertia weight, w; learning factors c_1 and c_2; number of variable skips, LNSsize; and M.

Generate the particle position and velocity:

$X_{pindex}^{iter=1}$ (representing the pindex particle at the first iteration for $1 \leq pindex \leq psize$) is expressed as follows: (v_1, v_2, . . . , v_i, . . . , v_N) there are N dimensions; for $1 \leq i \leq N$, v_i can be 0 (variable combination without variable i) or 1 (variable combination with variable i), and psize particles are chosen randomly as $X_{pindex}^{iter=1}$ (the position). The velocity $V_{pindex}^{iter=1}$ is randomly selected from the range $U[-V_{max}, V_{max}]$, and V_{max} is set according to 15% of the range of variables in each dimension [33].

4.2.2. Calculate the Particle Adaptation Values

The variables in the specific particle X_{pindex}^{iter} (for $1 \leq iter \leq isize$, $1 \leq pindex \leq psize$) are used as the inputs for long short-term memory (LSTM). After training and testing from real data, the root mean square error (RMSE) for X_{pindex}^{iter} is calculated and adopted as the particle adaptation value.

4.2.3. Execute LNS for the New Iteration

M particles are randomly selected from the new iteration (generation) of psize particles. LNSsize variables are arbitrarily selected for diversity for each selected particle in the N-dimensional variables. The value of the relevant variable is changed to 1 if it is 0 or to 0 if it is 1. We update these variables to generate a new solution, X. If the adaptation value of X is better than X_{pindex}^{iter}, then replace X_{pindex}^{iter}(X_{pindex}^{iter} = X), and if the adaptation value of X is worse than X_{pindex}^{iter}, then replace X_{pindex}^{iter}(X_{pindex}^{iter} = X) with the probability of $e^{\frac{f(X)-f(X_{pindex}^{iter})}{iter}}$.

4.2.4. Update Pbest and Gbest

We determine whether the adaptation value of each particle X_{pindex}^{iter} (for $1 \leq pindex \leq psize$) in iteration (generation) iter is better than $Pbest_{pindex}$ (set $Pbest_{pindex}$ = X_{pindex}^{iter} if it is the first generation of particles). If it is, then replace $Pbest_{pindex}$; thereafter, determine whether the Gbest update condition is met (if it is the first generation of particles, then set Gbest = best solution for all first-generation particles). If the particle is inferior to Gbest, then Gbest remains unchanged; if the particle is not inferior to Gbest, then the particle replaces Gbest, the particle updates the velocity and position, and LNS is performed for the particle (see Execute LNS for the New Iteration).

Has iter reached the default value (=isize)?

Has iter reached the default value of isize? If it has, (1) stop; otherwise, (2) update the particle velocity and position, (3) execute LNS for Gbest, (4) calculate iter = iter + 1, and (5) go back to calculate the particle adaptation values.

Update the particle position and velocity

Update the particle position X_{pindex}^{iter} and velocity V_{pindex}^{iter} according to the current position and velocity of each particle in the iteration (see Equations (7) and (8)) and check whether the combination of variables is out of range (0 and 1 for each variable). If the velocity is out of range, then the out-of-range velocity value is expressed as the maximum (out of maximum) or minimum (out of minimum) of the range. If the particle position (variable) out of X_{pindex}^{iter} has a non-integer variable (between 0 and 1), the upper limit (1) or lower limit (0) is used according to the nearest-distance principle:

$$V_{pindex}^{iter+1} = w \times V_{pindex}^{iter} + c_1 \times rand_1 \times \left(Pbest_{pindex} - X_{pindex}^{iter}\right) + c_2 \times rand_2 \times \left(Gbest - X_{pindex}^{iter}\right) \quad (7)$$

$$X_{pindex}^{iter+1} = X_{pindex}^{iter} + V_{pindex}^{iter+1} \quad (8)$$

4.2.5. Set the Related Parameters

Related parameters: number of particles, psize; number of iterations, isize; inertia weight, w; learning factors, c_1 and c_2; LNSsize; and M. The study carried out by Rabbani

et al. [33] was used to set the inertia weight w = 0.975 and the learning factors c_1 = 2 and c_2 = 2. Other values were determined experimentally based on the minimal RMSE criterion. The number of particles (psize) is tried from 50 to 150 (50, 100, and 150); the number of iterations (isize) is attempted from 100 to 300 (100, 200, and 300); LNSsize is attempted from 3 to 7 (3, 5, and 7); and M is attempted from 10 to 30 (10, 20, 30). The following hyperparameters of the LSTM prediction model are used in this paper after the experiments: activation function: rectified linear units (Relu), optimizer: Adam, number of hidden layers: one, hidden nodes: 30, epoch: 50, batch size: 10, learning rate: 0.001, and dropout rate: 0.4. The following values are determined after the experiments: psize = 50, isize = 200, LNSsize = 5, and M = 10.

4.3. The Data for the Crop Harvest Time Prediction Model

Bok choy, one of the most popular and important vegetables in Taiwan, was selected to verify the proposed model. The relevant data were mainly obtained from the production records of a large fruit and vegetable production cooperative in Yunlin County (Taiwan) for the past few years, together with the public climate-related database of the nearby Central Weather Bureau meteorological station. The proposed prediction model was trained with 10,000 data items, and 5025 data items were tested (some summary data samples from the production records and the public climate-related database mentioned above are listed in Table 2; all data samples are available upon request).

Table 2. Some summary data samples for model training and testing.

Date	Cumulative Accumulated Temperature	Accumulated Tempera-ture	Accumulated Sunshine Hours	Accumulated Total Sky Radiation	Accumulated Radiation	Accumulated Rainfall	Cumulative Precipitation Hours	Average Humidity	Average Wind Speed	Harvest Time
28 June 2019	333.2	20.4	11.1	23.5	3.0	0	0	71.8	305.8	14
29 June 2019	353.6	20.4	9.1	20.2	2.6	0	0	70.7	267.5	13
30 June 2019	373.5	19.9	7.1	16.9	2.5	0	1.4	74.8	195	12
1 July 2019	394.0	20.5	9.4	20.4	3.0	0	0	71.1	264.6	11
2 July 2019	413.3	19.3	2.6	13.8	2.3	0	1.4	72.8	315.8	10
3 July 2019	429.9	16.6	0	6.6	1.0	0	16.4	81.7	215	9
4 July 2019	449.4	19.5	4.9	16.3	2.4	0	0.2	68.3	226.7	8

5. Results

In this paper, three sets of model validation were designed: the recurrent neural network (RNN), long short-term memory without feature selection (LSTM), and long short-term memory with feature selection (LSTMFS, proposed in this paper and mentioned in Sections 4.1 and 4.2). The system environment of the experimental platform consisted of an Intel® Core™ i7-8700 CPU at 3.20GHz with 16GB RAM, and all validation programs and systems were built using Python 3.9.

Table 3 shows that the accuracy of LSTMFS (RMSE = 0.199; MAPE = 4.84%) is better than that of LSTM (RMSE = 0.565; MAPE = 15.92%) and RNN (RMSE = 1.327; MAPE = 28.89%). The best variable combination found by LSTMFS includes (1) cumulative accumulated temperature one day ago, (2) cumulative accumulated temperature two days ago, (3) cumulative accumulated temperature three days ago, (4) accumulative sunshine hours one day ago, (5) accumulative sunshine hours two days ago, (6) accumulative sunshine hours three days ago, (7) accumulated total sky radiation one day ago, (8) accumulated total sky radiation two days ago, and (9) accumulated total sky radiation three days ago.

Table 3. Comparison of prediction models.

Model	RMSE [1]	MAPE [2]
RNN	1.327	28.89%
LSTM	0.565	15.92%
LSTMFS	0.199	4.84%

[1] RMSE $= \sqrt{\frac{1}{K}\sum_{i=1}^{N}(h_i{}^* - h_i)^2}$, K: the number of testing data items, $h_i{}^*$: actual value, h_i: predicted value.

[2] MAPE $= \frac{1}{K}\sum_{i=1}^{N}\left|\frac{h_i{}^* - h_i}{h_i{}^*}\right|$, K: the number of testing data items, $h_i{}^*$: actual value, h_i : predicted value.

Table 4 shows the results of the RMSE prediction error analysis for LSTM and LSTMFS. The analysis result is significant (F = 949,017.2, significance < 0.001), and, therefore, the null hypothesis is rejected; i.e., the RMSE of LSTMFS is significantly lower than that of LSTM. LSTMFS is significantly better than LSTM.

Table 4. An ANOVA analysis for LSTM and LSTMFS.

	Sum of Squares	Degree of Freedom	Mean Square	F	Significance
Model	335.552	1	335.552	949,017.2	0.000
Error	3.553	10,048	0.000		
Total	339.104	10,049			

In this paper, we investigated whether the prediction model LSTMFS improves in accuracy as the prediction time to the harvesting period decreases from four weeks to three weeks, two weeks, and one week before harvesting. Table 5 shows the accuracy of the model's prediction regarding the prediction time to the harvesting period. It was found that the nearer the harvest time, the better the prediction accuracy. The RMSE values for the prediction times of one week before harvesting, two weeks before harvesting, three weeks before harvesting, and four weeks before harvesting, are 0.165, 0.185, 0.205, and 0.222, respectively.

Table 5. Accuracy of the prediction time to harvesting period.

Prediction Time to the Harvesting Period	RMSE
Four weeks	0.222
Three weeks	0.205
Two weeks	0.185
One week	0.165

Table 6 shows an ANOVA analysis for different prediction times to the harvesting period, in which the results are significant (F = 21,114.62, significance < 0.001), thus rejecting the null hypothesis that the four datasets are significantly different and indeed significantly different at different times. In addition, according to the pair comparison, the analysis results are significant for all pairs (Table 7). (1) For the one-week prediction time and the two-week prediction time, mean difference (I − J) = −0.019916745, standard error = 0.0002414928, and significance < 0.001. Therefore, the null hypothesis is rejected; i.e., the RMSE of the one-week prediction time is significantly lower than that of the two-week prediction time. The one-week prediction time is significantly better than the two-week prediction time. (2) For the one-week prediction time and the three-week prediction time, mean difference (I − J) = −0.039959544, standard error = 0.0002413134, and significance < 0.001. Therefore, the null hypothesis is rejected; i.e., the RMSE of the one-week prediction time is significantly lower than that of the three-week prediction time. The one-week prediction time is significantly better than the three-week prediction time. (3) For the one-week prediction time and the four-week prediction time, mean difference (I − J) = −0.057357473,

standard error = 0.0002415528, and significance < 0.001. Therefore, the null hypothesis is rejected; i.e., the RMSE of the one-week prediction time is significantly lower than that of the four-week prediction time. The one-week prediction time is significantly better than the four-week prediction time. (4) For the two-week prediction time and the three-week prediction time, mean difference (I − J) = −0.020042799, standard error = 0.0002411932, and significance < 0.001. Therefore, the null hypothesis is rejected; i.e., the RMSE of the two-week prediction time is significantly lower than that of the three-week prediction time. The two-week prediction time is significantly better than the three-week prediction time. (5) For the two-week prediction time and the four-week prediction time, mean difference (I − J) = −0.037440728, standard error = 0.0002414327, and significance < 0.001. Therefore, the null hypothesis is rejected; i.e., the RMSE of the two-week prediction time is significantly lower than that of the four-week prediction time. The two-week prediction time is significantly better than the four-week prediction time. (6) For the three-week prediction time and the four-week prediction time, mean difference (I − J) = −0.017397929, standard error = 0.0002412532, and significance < 0.001. Therefore, the null hypothesis is rejected; i.e., the RMSE of the three-week prediction time is significantly lower than that of the four-week prediction time. The three-week prediction time is significantly better than the four-week prediction time.

Table 6. An ANOVA analysis for different prediction times to the harvesting period.

	Sum of Squares	Degree of Freedom	Mean Square	F	Significance
Prediction time	1.856	3	0.619	21,114.62	0.000
Error	0.118	4020	0.000		
Total	1.974	4023			

Table 7. Multiple comparisons of the prediction time to the harvesting period.

Week (I)	Week (J)	Mean Difference (I − J)	Standard Error	Significance
1	2	−0.019916745 *	0.0002414928	0.000
	3	−0.039959544 *	0.0002413134	0.000
	4	−0.057357473 *	0.0002415528	0.000
2	1	0.019916745 *	0.0002414928	0.000
	3	−0.020042799 *	0.0002411932	0.000
	4	−0.037440728 *	0.0002414327	0.000
3	1	0.039959544 *	0.0002413134	0.000
	2	0.020042799 *	0.0002411932	0.000
	4	−0.017397929 *	0.0002412532	0.000
4	1	0.057357473 *	0.0002415528	0.000
	2	0.037440728 *	0.0002414327	0.000
	3	0.017397929 *	0.0002412532	0.000

Note: 1: previous one week, 2: previous two weeks, 3: previous three weeks, 4: previous four weeks. *: the mean difference is significant at the 0.05 level.

6. Discussion

The proposed model can predict the harvesting period accurately (RMSE = 0.199; MAPE = 4.84%) so as to help achieve a balance between production and sales in the sustainable supply chain and reduce resource waste for better sustainability [1,11,34].The feature selection (variable selection) method was adopted to select the features (variables) that can effectively represent input features (variables) of the model and reduce the complexity of the model, and better prediction results were obtained (LSTMFS is significantly better than LSTM. Please refer to Tables 3 and 4 for details.) [18]. Furthermore, it was observed that LSTM (RMSE = 0.565; MAPE = 15.92%) is a better prediction model than the RNN (RMSE = 1.327; MAPE = 28.89%). The results are the same as those in previous studies [13,14]. Moreover, since the harvest time is nearer in the case of LSTMFS, this model can obtain more related data and learn more from the data, and the prediction accuracy

is better (Please refer to Tables 5–7 for details). In addition, better sustainability can be achieved since the accuracy of the prediction model is improved [1,11].

Table 8 shows a comparison of crop harvest time prediction in different studies. Compared with other existing studies, the proposed model, LSTMFS, which integrates a hybrid search for feature selection and a long short-term memory (LSTM) model, proved to be a novel and effective method. The RMSE (=0.199) for LSTMFS is much better than those reported by [1] (RMSE = 2.58), [2] (RMSE = 5.5), and [5] (RMSE = 0.5176). The MAPE (=4.84%) for LSTMFS is better than that reported by [6] (MAPE = 6%).

Table 8. Comparison of harvest time prediction in different studies.

Source	Crop	Prediction Method	Feature Selection Method	Metric for Methods
This paper	Bok choy	LSTM	A hybrid search (PSO and LNS)	RMSE (=0.199), MAPE (=4.84%)
[1]	Apple	RNN	Empirical	RMSE (=2.58)
[2]	Barley, wheat	Statistical method	No	RMSE (=5.5 for both)
[5]	Lettuce	ANN	Empirical	RMSE (=0.5176)
[6]	Banana	ANN	Empirical	MAPE (=6%)
[10]	Broccoli	Statistical method	No	RMSD, RMAE
[11]	Tomato	CNN	No	Accuracy

7. Conclusions

This paper develops a novel crop harvest time prediction model, LSTMFS, which integrates a hybrid search for feature selection and a long short-term memory (LSTM) model based on real production and climate-related data, in order to accurately predict harvest time and reduce resource waste for better sustainability. Based on the results, LSTMFS is significantly better than long short-term memory (LSTM) and recurrent neural networks (RNNs). Moreover, the nearer the harvest time, the better the prediction accuracy. In addition, compared with other existing studies, the proposed model, LSTMFS, proves to be an effective method.

In a future research direction, we hope to combine different sensors (such as soil and ultraviolet light) to collect more distinguishing features that affect the harvesting period, which will be helpful in improving the accuracy of the model prediction. In addition to harvest time, other important growth indicators (such as crop harvest rate, crop growth rate, and yield) can also be investigated.

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

Funding: This research was financially supported by grants of the National Science and Technology Council, Taiwan (111-2637-E-020-004).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

Abbreviations

ACO	Ant colony optimization algorithm
AI	Artificial intelligence
ANN	Artificial neural networks
ANOVA	Analysis of variance
CNN	Convolutional neural networks
GA	Genetic algorithms
LNS	Large neighborhood search
LSTM	Long short-term memory
LSTMFS	Long short-term memory with feature selection
MAPE	Mean absolute percentage error
ML	Machine learning
MLP	Multilayer perceptron
PSO	Particle swarm algorithms
RMSD	Root mean square deviation
RMSE	Root mean square error
RNN	Recurrent neural networks
SA	Simulated annealing
TS	Taboo search
VNS	Variable neighborhood search

References

1. Boechel, T.; Policarpo, L.M.; Ramos, G.d.O.; da Rosa Righi, R.; Singh, D. Prediction of harvest time of apple trees: An RNN-based approach. *Algorithms* **2022**, *15*, 95. [CrossRef]
2. Pullens, J.W.M.; Sørensen, C.A.G.; Olesen, J.E. Temperature-based prediction of harvest date in winter and spring cereals as a basis for assessing viability for growing cover crops. *Field Crops Res.* **2021**, *264*, 108085. [CrossRef]
3. Hamilton, J.D. *Time Series Analysis*; Princeton University Press: Princeton, NJ, USA, 1994; pp. 105–175.
4. Munoz, C.; Ávila, J.; Salvo, S.; Huircán, J.I. Prediction of harvest start date in highbush blueberry using time series regression models with correlated errors. *Sci. Hortic.* **2012**, *138*, 165–170. [CrossRef]
5. Chang, C.L.; Chung, S.C.; Fu, W.L.; Huang, C.C. Artificial intelligence approaches to predict growth, harvest day, and quality of lettuce (*Lactuca sativa* L.) in a IoT-enabled greenhouse system. *Biosyst. Eng.* **2021**, *212*, 77–105. [CrossRef]
6. de Souza, A.V.; Neto, A.B.; Piazentin, J.C.; Junior, B.J.D.; Gomes, E.P.; Bonini, C.D.S.B.; Putti, F.F. Artificial neural network modelling in the prediction of bananas' harvest. *Sci. Hortic.* **2019**, *257*, 108724. [CrossRef]
7. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
8. Karevan, Z.; Suykens, J.A.K. Transductive LSTM for time-series prediction: An application to weather forecasting. *Neural Netw.* **2020**, *125*, 1–9. [CrossRef] [PubMed]
9. Taha, A.; Cosgrave, B.; Mckeever, S. Using feature selection with machine learning for generation of insurance insights. *Appl. Sci.* **2022**, *12*, 3209. [CrossRef]
10. Lindemann-Zutz, K.; Fricke, A.; Stützel, H. Prediction of time to harvest and its variability of broccoli (*Brassica oleracea* var. italica) part II. Growth model description, parameterisation and field evaluation. *Sci. Hortic.* **2016**, *200*, 151–160. [CrossRef]
11. Minagawa, D.; Kim, J. Prediction of Harvest Time of Tomato Using Mask R-CNN. *AgriEngineering* **2022**, *4*, 24. [CrossRef]
12. Zhang, J.; Wang, P.; Yan, R.; Gao, R.X. Long short-term memory for machine remaining life prediction. *J. Manuf. Syst.* **2018**, *48*, 78–86. [CrossRef]
13. Kumari, P.; Toshniwal, D. Deep learning models for solar irradiance forecasting: A comprehensive review. *J. Clean. Prod.* **2021**, *318*, 128566. [CrossRef]
14. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D.* **2020**, *404*, 132306. [CrossRef]
15. Sagheer, A.; Kotb, M. Time series forecasting of petroleum production using deep LSTM recurrent networks. *Plant Physiol.* **2019**, *100*, 2106–2108. [CrossRef]
16. Yadava, A.; Jhaa, C.K.; Sharan, A. Optimizing LSTM for time series prediction in Indian stock market. *Procedia Comput. Sci.* **2020**, *167*, 2091–2100. [CrossRef]
17. Chimmula, V.K.R.; Zhang, L. Time series forecasting of COVID-19 transmission in Canada using LSTM network. *Chaos Solit. Fractals* **2020**, *135*, 109864. [CrossRef]
18. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr Eng.* **2014**, *40*, 16–28. [CrossRef]
19. Hansen, P.; Mladenović, N. Variable neighborhood search: Methods and applications. *Ann. Oper. Res.* **2010**, *175*, 367–407. [CrossRef]
20. Costa, H.; Galvao, L.R.; Merschmann, L.H.C.; Souz, M.J.F. A VNS algorithm for feature selection in hierarchical classification context. *Electron. Notes Discret. Math.* **2018**, *66*, 79–86. [CrossRef]

21. Wang, Y.; Lü, Z.; Su, Z. A two-phase intensification tabu search algorithm for the maximum min-sum dispersion problem. *Comput. Oper. Res.* **2021**, *135*, 105427. [CrossRef]
22. Üstünkar, G.; Özöğür-Akyüz, S.; Weber, G.W.; Friedrich, C.M.; Aydın, S.Y. Selection of representative SNP sets for genome-wide association studies: A metaheuristic approach. *Optim. Lett.* **2012**, *6*, 1207–1218. [CrossRef]
23. Garcia-Torres, M.; Gomez-Vela, F.; Melian, B.; Moreno-Vega, J.M. High-dimensional feature selection via feature grouping: A variable neighborhood search approach. *Inf. Sci.* **2016**, *326*, 102–118. [CrossRef]
24. Rostami, M.; Berahmand, K.; Nasiri, E.; Forouzandeh, S. Review of swarm intelligence-based feature selection methods. *Eng. Appl. Artif. Intel.* **2021**, *100*, 104210. [CrossRef]
25. Moradi, P.; Gholampour, M. A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Appl. Soft Comput.* **2016**, *43*, 117–130. [CrossRef]
26. Esfandiarpour-Boroujeni, I.; Karimi, E.; Shirani, H.; Esmaeilizadeh, M.; Mosleh, Z. Yield prediction of apricot using a hybrid particle swarm optimization-imperialist competitive algorithm-support vector regression (PSO-ICA-SVR) method. *Sci. Hortic.* **2019**, *257*, 108756. [CrossRef]
27. Li, W.; Becker, D.M. Day-ahead electricity price prediction applying hybrid models of LSTM-based deep learning methods and feature selection algorithms under consideration of market coupling. *Energy* **2021**, *237*, 121543. [CrossRef]
28. Hatfield, J.L.; Prueger, J.H. Temperature extremes: Effect on plant growth and development. *Weather Clim. Extrem.* **2015**, *10*, 4–10. [CrossRef]
29. Punia, H.; Tokas, J.; Malik, A.; Kumar, S. Solar radiation and nitrogen use efficiency for sustainable agriculture. In *Resources Use Efficiency in Agriculture*; Kumar, S., Meena, R.S., Jhariya, M.K., Eds.; Springer: Singapore, 2020; pp. 177–212.
30. Ndamani, F.; Watanabe, T. Influences of rainfall on crop production and suggestions for adaptation. *Int. J. Agric. Sci.* **2015**, *5*, 367–374.
31. Hirai, G.; Okumura, T.; Takeuchi, S.; Tanaka, O.; Chujo, H. Studies on the effect of the relative humidity of the atmosphere on the growth and physiology of rice plants. *Plant Prod. Sci.* **2000**, *3*, 129–133. [CrossRef]
32. Gardiner, B.; Berry, P.; Moulia, B. Review: Wind impacts on plant growth, mechanics and damage. *Plant Sci.* **2016**, *245*, 94–118. [CrossRef]
33. Rabbani, M.; Bajestani, M.A.; Khoshkhou, G.B. A multi-objective particle swarm optimization for project selection problem. *Expert Syst. Appl.* **2011**, *37*, 315–321. [CrossRef]
34. Alkahtani, M. Supply chain management optimization and prediction model based on projected stochastic gradient. *Sustainability* **2022**, *14*, 3486. [CrossRef]