A Strategy of Assessing Climate Factors' Influence for Agriculture Output

Chin-Hung Kuan[1], Yungho Leu[2], and Chien-Pang Lee[3*]
[1,2] Department of Information Management, National Taiwan University of Science and Technology, Taipei City 106, Taiwan
[e-mail : D10509101@mail.ntust.edu.tw, yhl@mail.ntust.edu.tw]
[3] Department of Maritime Information and Technology, National Kaohsiung University of Science and Technology, Kaohsiung City 805, Taiwan
[e-mail : cplee@nkust.edu.tw]
[*]Corresponding author : Chien-Pang Lee

Received June 18, 2021; revised July 22, 2021; revised December 24, 2021; revised March 24, 2022; accepted April 30, 2022; published May 31, 2022

Abstract

Due to the Internet of Things popularity, many agricultural data are collected by sensors automatically. The abundance of agricultural data makes precise prediction of rice yield possible. Because the climate factors have an essential effect on the rice yield, we considered the climate factors in the prediction model. Accordingly, this paper proposes a machine learning model for rice yield prediction in Taiwan, including the genetic algorithm and support vector regression model. The dataset of this study includes the meteorological data from the Central Weather Bureau and rice yield of Taiwan from 2003 to 2019. The experimental results show the performance of the proposed model is nearly 30% better than MARS, RF, ANN, and SVR models. The most important climate factors affecting the rice yield are the total sunshine hours, the number of rainfall days, and the temperature.The proposed model also offers three advantages: (a) the proposed model can be used in different geographical regions with high prediction accuracies; (b) the proposed model has a high explanatory ability because it could select the important climate factors which affect rice yield; (c) the proposed model is more suitable for predicting rice yield because it provides higher reliability and stability for predicting. The proposed model can assist the government in making sustainable agricultural policies.

Keywords: Climate factor, Support vector regression, Genetic algorithm, Rice yield, Machine learning

1. Introduction

The Food and Agriculture Organization of the United Nations (FAO) reported the food and cereals price index, which had increased 200% and 300% approximately, respectively, in 2009 [1]. Furthermore, because cereals are the staple food in many countries, the price increase might have exacerbated 100 million people [2]. The stocks of world cereal in 2008 were the lowest for the last 25 years [3]. Accordingly, the food crisis seems to be imminent. Although the increase tended for the food price index to be flat, it is still 150% higher than in 2001, according to the new FAO report in 2019 [4], as shown in **Table 1**. Because of this reason, the food crisis is still a severe subject now.

Because the agriculture industry plays a vital role in the stable operation of national economies [5-7], many countries require an adequate basis to formulate agricultural policies to alleviate the potential effects of food crises [8]. However, agricultural data is highly volatile due to climate change [9]. Additionally, many uncertainties are in agriculture, which imposes an additional burden on farmers [10]. These reasons explain the difficulty of agricultural prediction. However, an excellent accurate agricultural prediction could be used to reduce uncertainties risks. Thus, an excellent agricultural prediction could be used to assist the government in managing and marking agricultural policy [2].

Table 1. FAO food price index [4]

Period	Food price index	Meat	Dairy	Cereals	Vegetables	Sugar
2001	94.6	100.1	105.5	86.8	67.2	122.6
2002	89.6	59.9	80.9	93.7	87.4	97.8
2003	97.7	95.9	95.6	99.2	100.6	100.6
2004	112.7	114.2	123.5	107.1	111.9	101.7
2005	118.0	123.7	135.2	101.3	102.7	140.3
2006	127.2	120.9	129.7	118.9	112.7	209.6
2007	161.4	130.8	219.1	163.4	172.0	143.0
2008	201.4	160.7	223.1	232.1	227.1	181.6
2009	160.3	141.3	148.6	170.2	152.8	257.3
2010	188.0	158.3	206.6	179.2	197.4	302.0
2011	229.9	183.3	229.5	240.9	254.5	368.9
2012	213.3	182.0	193.6	236.1	223.9	305.7
2013	209.8	184.1	242.7	219.3	193.0	251.0
2014	201.8	198.3	224.1	191.9	181.1	241.2
2015	164.0	168.1	160.3	162.4	147.0	190.7
2016	161.5	156.2	153.8	146.9	163.8	256.0
2017	174.6	170.1	202.2	151.6	168.8	227.3
2018	168.4	166.3	192.9	165.3	144.0	177.5
2019	171.4	175.6	198.7	164.3	135.2	180.3

Note that The FAO Food Price Index is a measure of the monthly change in international prices of a basket of food commodities. It consists of the average of five commodity group price indices, weighted with each of the groups' average export shares for 2002-2004 [4].

To achieve an accurate agricultural prediction, many countries have developed precision agriculture to manage the farm or make the agricultural policy [11]. Before developing precision agriculture, the most important issue is collecting data to support the decision from the analysis. However, many factors may affect agriculture output, especially climatic factors. Accordingly, how to collect kinds of climatic data for precision agriculture is also a difficult issue. Fortunately, the kinds of climatic data could be collected simultaneously by IoT

(Internet of Things).

Because only collecting a larger number of data or kinds of data is not enough to decide on precision agriculture, the second stage of precision agriculture must be constructing an accurate agricultural prediction model. We then used the model to find important factors and their critical values to help researchers make the right decisions.

As described above, accurate agricultural prediction models can assist in developing precision agriculture and help formulate appropriate agricultural policies [5]. Therefore, many researchers focus on agriculture research issues [2, 7, 12-14]. In the past, statistical methods are widely applied in agriculture research issues. Although statistical methods could be used to solve many agriculture research issues, the statistical assumptions might not hold [13].

Taiwan also pays attention to agriculture research issues because it is a flourishing agriculture region. Taiwan is an island near Mainland China, as shown in **Fig. 1**, and its location is across the subtropical and tropical regions. Accordingly, the agricultural products of Taiwan are diverse and abundant. Although Taiwan's agricultural industry has developed very well, it still suffers from many agricultural risks every year. One of the risks is the climate factor. For example, typhoons or heavy rains often affect the harvests of rice. Accordingly, this paper proposes a hybrid model to confirm the influence of climate factors in predicting rice yield. The model is based on machine learning techniques without statistical assumptions and explains the results more clearly and directly.

The proposed hybrid model predicts rice yield to reduce food crises based on climate data. The model includes the genetic algorithm (GA) and the support vector regression (SVR) model. Because the SVR model is nonlinear, this paper uses it as the primary prediction method without statistical assumptions. Furthermore, the GA in the model is used to select a suitable candidate set of variables from the climate variables to enhance the prediction performance. Furthermore, due to the location of Taiwan across the subtropical and tropical regions, the characteristics of climate factors are different from different geographical regions. This paper also compares the difference in the influence of climate factors for other geographical regions.

The remainder of this paper is organized as follows. Section 2 introduces the related works, and Section 3 explains the proposed model. Subsequently, we report and discuss the experiment results in Section 4 and conclude the paper in Section 5.

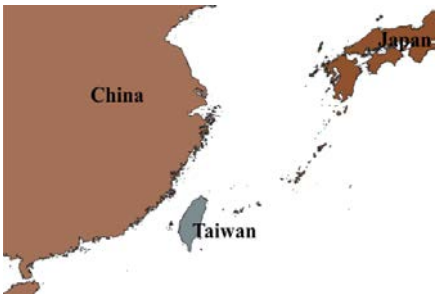

Fig. 1. The location of Taiwan.

2. Related Works

Section 2 includes six subsections. Section 2.1 arranges the literature review of agricultural prediction based on climate data and discusses the reason for constructing a prediction model in this study. Section 2.2 and Section 2.3 introduce the main methods in the proposed model. Section 2.4 to Section 2.7 introduce the main comparison models and the performance measures.

2.1 Literature Review of Agricultural Prediction

By reviewing the literature, an unstable food supply would affect the stable operation of national economies and the survival of the country [5-7]. Accordingly, although agriculture is not the primary industry in many countries, many researchers still study agricultural issues to prevent the potential effects of food crises [8]. In past agricultural research, experimental designs, one of the statistical methods, were widely used to discuss effect factors or treatments to increase the agricultural quantity [15, 16]. With the rapid development of computers, many data mining or machine learning models were successfully applied in agriculture issues [2, 5, 17].

Due to climate change, many uncertainties in agriculture have resulted in highly volatile agricultural data [9]. Therefore, many researchers proposed or applied data mining or machine learning models based on climate data to predict agricultural issues [17-19]. Kaul et al. [17] used rainfall information to construct the artificial neural networks (ANN) model for predicting corn and soybean yield. Tian et al. [19] successfully used five climate indices based on the support vector regression (SVR) for predicting the agricultural drought issue. Shin et al. [18] used a coupled global model and machine learning models, such as ANN models, to predict daily mean air temperatures for field-scale agricultural management.

Although the literature pointed out that climate factors are important for agriculture prediction models, many climate factors might have different important degrees in different months that resulted from the growth periods and might simultaneously affect agricultural issues. Based on this reason, this paper considers five climate factors in different months to predict rice yield.

2.2 Support Vector Regression

Initially, the support vector machine (SVM) model was proposed for classification problems in 1992. Now, the SVM is a well-known classifier [20, 21]. Subsequently, the SVM model was extended to solve nonlinear regression estimation issues in 1996 [22]. Accordingly, the extended SVM model is also called the SVR model. The SVR model was recently successfully applied in many agricultural research issues [2, 12, 13]. We briefly review the algorithm of the SVR model as follows. Firstly, x_i and y_i are assumed to be the ith input (independent) variables, and the ith corresponding target (dependent) variables, respectively, of a dataset (x_i, y_i) (i = 1, 2, …, l; $x_i \in R^d$; $y_i \in R$), where d denotes dimensions and R denotes the real number. Secondly, searching a function $f(x)$ that exhibits a deviation ε from the actual y_i for all training datasets and is as flat as possible [23, 24]. Subsequently, the function $f(x) = wx + b$ is assumed, where $w \in X,\ b \in R$ (X denotes the input space, and R denotes the real number). The function $f(x)$ can then be solved by using the following equations:

$$\min\left\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)\right\} \quad \text{s.t.}\begin{cases} y_i - wx_i - b \le \varepsilon \\ wx_i + b - y_i \le \varepsilon \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{1}$$

$$|\xi|_\varepsilon = \begin{cases} |\xi| - \varepsilon, & if \ |\xi| > \varepsilon \\ 0, & if \ |\xi| \le \varepsilon \end{cases} \tag{2}$$

In Eq. (1) and Eq. (2), C determines the tradeoff between the flatness of function $f(x)$ and the value up to which deviations are greater than ε are tolerated; ξ_i and ξ_i^* are positive slack

variables; $|\xi|_{\varepsilon}$ is the ε-insensitive loss function [2, 13, 23, 24]. To solve the optimization as mentioned above problem, they are translated into a Lagrange dual problem, as shown in Eq. (3), and the solution is derived according to Eq. (4). In Eq. (3) and Eq. (4), α_i and α_i^* are the Lagrange multipliers corresponding to ξ and ξ_i^*; $k(x_i, x_j)$ is the kernel function. Generally, the SVR model has three kinds of main kernel functions [25]. The first is the Gaussian radial basis function (RBF) kernel function, which includes two parameters (*Cost* and *Gamma*). The second is the polynomial kernel function, consisting of three parameters (*Cost*, *Gamma*, and *degree*). The third is the sigmoid kernel function, including two parameters (*Cost* and *Gamma*).

$$\max -\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)k(x_i,x_j)/2-\varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*)$$
$$\text{s.t}\begin{cases}\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0\\ 0\le\alpha_i,\alpha_i^*\le C\end{cases} \tag{3}$$

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)k(x_i,x_j)+b \tag{4}$$

2.3 Genetic Algorithm

There are many variable selection methods, such as statistical methods and optimal algorithms. This paper chooses the optimal algorithm as a variable selection method without considering the statistical assumptions described in Section 1. Many optimal algorithms are proposed or used for feature selection. The most used is the genetic algorithm (GA). Although GA is not the newest or the most efficient optimization algorithm, GA is a commonly used and simple optimization algorithm for variable selection in agricultural issues [5, 12, 13]. Accordingly, this paper integrated GA with the proposed model.

In the 1970s, John Holland first proposed GA [26]. Now, GA is a well-known algorithm and is widely used to search for exact or approximate solutions. GA searches for the optimal solution is based on imitating a biological system [12, 27]. Recently, many new algorithms have been proposed to modify the performance of the traditional GA, but their structures are almost based on the traditional GA. A traditional GA at least includes six steps, as shown in **Fig. 2**, to randomly generate many chromosomes (solutions) as a population and then calculate their fitness value. Subsequently, GA improves chromosomes through three primary operations: selection, crossover, and mutation operations. Finally, GA is confirmed to continue generating the next generation or terminating the algorithm to select the best solution.

2.4 Random forest regression

The random forest regression (RF) model is an ensemble learning algorithm for classification and prediction [28, 29]. When it is used on prediction issues, it is also called the random forest regression model. The RF model is a tree-based regression model based on constructing a multitude of decision regression trees [30]. Recently, the RF model is also a widely used algorithm in many research issues as it brings better performance, ease of implementation, and low computational cost [28, 30]. Accordingly, the RF model was selected as the first comparison model.

The RF model is a combination of decision trees [30]. This combination reduces the error in classification and regression research issues through bootstrap or bagging methods [31]. Accordingly, the central concept of the RF model is to minimize the error of the prediction issues into account for the decision trees included within the forest and the correlation among their predictions [31, 32]. The main parameters of the RF model are:
– m: the number of variables considered in each node.
– T: the number of trees in the forest.
– M_s: elements in a node required to perform a split.
– M_l: elements required to create a node.
– L: the maximum depth up to which a tree can grow.

The detailed procedure of the RF model could refer to the literature [28, 30, 31]. The RF model was implemented by the "random Forest" package of the *R* language in this paper.

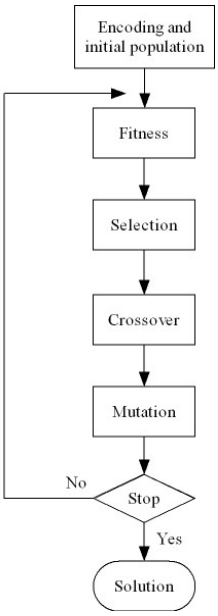

Fig. 2. The structure of the traditional genetic algorithm.

2.5 Artificial neural networks

In the 1950s, the artificial neural network (ANN) concept was first introduced. Many kinds of ANN models have been proposed recently, such as the multilayer perceptron neural network, the feedforward neural network, the backpropagation neural network, and the radial basis function neural network [13, 33-35].

As described in Section 2.1, the ANN model was chosen as the second comparison model because the ANN model has been successfully applied in agricultural or climatic issues in recent years [13, 17, 18, 33]. Although many kinds of ANN models are proposed, their basic frameworks are similar. The basic framework of the ANN model contains three layers, such as the input layer, the hidden layer, and the output layer, as shown in **Fig. 3**. The nodes in the input layer denote input variables, and the nodes in the hidden layer denote the activation procedure based on the chosen activation function of the ANN model. The node in the output layer denotes the rice yield in this paper. Note that the ANN model was implemented by the "nnet" package of the *R* language in this paper.

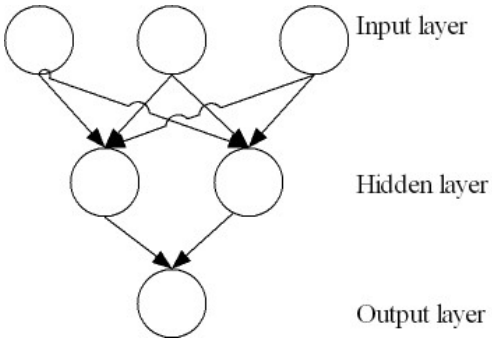

Fig. 3. The basic framework of artificial neural networks with three layers.

2.6 Multivariate Adaptive Regression Splines

The multivariate adaptive regression splines (MARS) model was first introduced in 1991 [36]. It is a non-parametric regression model and an extension of linear models that automatically nonlinearities and interactions between variables. Thus, the MARS model could be used for flexible regression modeling of the high-dimensional data.

Recently, The MARS model is also successfully applied in geographic or spatial issues [37, 38]. Because the analyzed dataset of this paper is high-dimensional and implied geographical information, the MARS model is used to compare the prediction performance with the proposed model. Note that the MARS model was implemented by the "mda" package of the *R* language in this paper.

2.7 Performance measures

Two performance measures, which are the MAE (mean absolute error) and the RMSE (root mean square error), are used to verify the prediction accuracy.

The MAE measures errors between the paired actual value and the predicted value. The MAE is defined in Eq. (5), P_t, and $\widehat{P}_t$ denotes the actual rice yield and the predicted rice yield in year *t*, respectively; *n* denotes the number of samples.

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|P_t - \widehat{P}_t\right| \tag{5}$$

The RMSE is also widely used to measure the performance in prediction research issues [24, 39]. A model with a lower RMSE denotes its prediction accuracy is better than the others. The equation of the RMSE is defined in Eq. (6).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(P_t - \widehat{P}_t)^2} \tag{6}$$

where P_t and $\widehat{P}_t$ denote the actual rice yield and the predicted rice yield in year *t*, respectively; *n* denotes the number of samples.

According to Eq. (5) and Eq. (6), a lower MAE or RMSE value indicates a more accurate forecasting power.

3. The proposed model

Because of the characteristics of agriculture data, Section 3.1 first introduces the used dataset

and its characteristics. Subsequency, the procedure of the proposed model is introduced in Section 3.2.

3.1 Dataset

The rice yield and the climate data from 2003 to 2019 are collected from the Annual Report of the Council of Agriculture and the Central Weather Bureau, respectively, in Taiwan. Because only 12 geographical regions have complete climate data and rice yield, we use the data of 12 regions to implement the proposed model. The collected data includes climate data, rice planting area for the first crop season, and rice yield for the first crop season from the 12 regions. The sample size of the collected data is 204 (17 years * 12 regions) samples.

The climate data includes five variables (sunshine hours, rainfall days, rainfall, average temperature, and average relative humidity), and each climate variable contains five months. **Table 2** gives an example for analyzing the Nantou region. In **Table 2**, X_{ij} denotes the ith climate variable and the jth month (j=1 for February, j=2 for March, j=3 for April, j=4 for May, j=5 for June; i=1 for sunshine hours, i=2 for rainfall days, i=3 for rainfall, i=4 for average temperature, and i=5 for average relative humidity). Furthermore, the AREA denotes the rice planting area for the first crop season; RY denotes the rice yield for the first crop season.

Table 2. The analyzed data in Nantou.

Region	Year	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	…	X_{55}	AREA	RY
Nantou	2003	173.2	130.0	97.0	136.4	125.4	…	85	2546	13.422
	2004	183.3	107.8	138.7	136.0	127.4	…	84	2661	13.634
	⋮	⋮	⋮	⋮	⋮	⋮	…	⋮	⋮	⋮
	2019	161.0	127.6	156.4	54.4	89.8	…	84	2343	12.683

Note: (a) the unit of sunshine hours is hours; the unit of rainfall days is day; (b) the unit of rainfall is cumulative mm; (c) the unit of average temperature is ºC; (d) the unit of average relative humidity is %; (e) the unit of AREA is hectare, and (f) the unit of RY is thousand tons.

3.2 The proposed model

The proposed model includes four phases. The first phase is the reorganizing data phase to reorganize data into a suitable form for analysis. The second phase is the splitting dataset phase to split data into the training dataset and the test dataset for building the training model and verifying the performance of the model, respectively. The third phase is the training phase to select a candidate set of variables from the training dataset. The final phase is the prediction phase to verify the proposed model's performance and predict the future rice yield. The detailed procedure is shown as follows.

The main goal of this paper is to predict the rice yield of next year yield by using the climate data of the current year and the rice planting area. Accordingly, the initially collected dataset must be reorganized into the proper prediction format. We replace the rice yield of the current year for the first crop season with the rice yield of next year for the first crop season in the first phase. For example, we used the climate data and the rice planting area for the first crop season in 2003 to predict the rice yield for the first crop season in 2004. Accordingly, the sample size of the reorganized dataset is only 192 samples.

In the second phase, we randomly split the dataset into a training dataset (70%) and a test dataset (30%) from each region and perform this procedure 100 times to generate 100 sets of training and test datasets for each region to enhance the randomness of data. Subsequently, each region's 100 sets of training and test datasets are used to build models and verify their performance.

The third phase is the training phase. Phase 3 includes two steps (GA and SVR model). GA selects a suitable candidate set of variables for the SVR model to predict the next year's rice yield more accurately. In this phase, the fitness function of GA is defined as Eq. (7),

$$fitness = \frac{RMSE + MAE}{2} \tag{7}$$

In Eq. (7), RMSE denotes the training root mean square error of the SVR model, and MAE denotes the training mean absolute error of the SVR model. The input (independent) variables include all climate data and the rice planting area for the first crop season, and the output (dependent) variable is the next year's rice yield for the first crop season.

The exhaustive procedures of Phase 3 are described as follows.

- Encoding chromosomes: we use binary encoding to generate chromosomes in this paper. Initially, each chromosome includes 26 genes (bits) because the dataset contains 26 input variables (five climate variables contain five months and a rice planting area variable). Generally, when a gene (bit) of a chromosome is encoded as 1, the corresponding variable is selected to train the SVR model.
- Calculating fitness: after generating chromosomes, all chromosomes must be calculated as the fitness values. A chromosome with a better fitness value denotes that its set of the selected variables is more suitable to the SVR model. In the proposed model, we use the definition of Eq. (7) to compute the fitness value of each chromosome. Namely, a chromosome with the lower RMSE and MAE denotes that it has the better fitness value.
- Performing genetic operations: after calculating the fitness values of chromosomes, three genetic operations, including roulette selection, one-point crossover, and uniform mutation, are used to generate the next generation.
- Stop algorithm: there are two termination conditions in this paper, according to the literature [40, 41], the first termination condition is when the generations are greater than 2,000 generations, and the second termination condition is when the best chromosomes of the last 200 generations are the same.

Due to the characteristic of the GA, the best chromosome of each GA procedure might not be the same. According to the literature concepts [40], we perform the GA procedure 100 times to generate 100 best chromosomes. Then we calculate the selected frequency of each gene (bit) from the 100 chromosomes. When the selected frequency of a gene (bit) is greater than 50 times, that gene's corresponding variable is selected for the final phase.

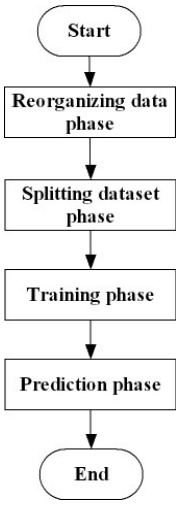

Fig. 4. The flowchart of the proposed model.

The final phase is the prediction phase. We use the selected variables from Phase 3 to build the proposed model and then use the build model to predict the remainder of the test data. Finally, the performance measures calculate the prediction accuracy of the proposed model. The flowchart of the proposed model is shown in **Fig. 4**.

In this paper, we used the "e1071" package of the *R* language to implement the SVR model, and then the parameters, such as Cost and Gamma, of the SVR model were automatically determined by the "tune.svm" function of "e1071" package. For the GA, all procedures of the algorithm were implemented by the authors.

4. Results

4.1 Data reorganized

According to Section 3.2, the initially collected dataset must be reorganized for the following prediction procedure. A part of the reorganized dataset is shown in **Table 3**. The main difference between **Table 2** and **Table 3** is that **Table 3** replaces **RY (**the rice yield for the first crop season in the same year**)** with **NRY (the next year's rice yield)**. To simplify, the value of NRY (13.634) in the first row of **Table 3** denotes the rice yield in 2004. Accordingly, the sample size has changed from 204 to 192 after the reorganizing data phase.

Table 3. An example of the reorganized data.

Region	Year	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	…	X_{55}	AREA	NRY
Nantou	2003	173.2	130.0	97.0	136.4	125.4	…	85	2546	13.634
	2004	183.3	107.8	138.7	136.0	127.4	…	84	2661	13.670
	⋮	⋮	⋮	⋮	⋮	⋮	…	⋮	⋮	⋮
	2018	161.0	127.6	156.4	54.4	89.8	…	84	2343	12.683

4.2 Splitting the dataset

After the reorganizing data phase, we split the dataset into training and test dataset. Because this paper considers comparing the performance of the models in the whole region (Taiwan) and individual regions, we first split the dataset into 13 different groups (Taiwan and the other 12 regions (cities or countries). Subsequently, for each group, we randomly select 70% of the data from the group as the training dataset and the remainder data are as the test dataset. To enhance the randomness for selecting the training and the test dataset in each group, we perform this procedure in each group 100 times to generate 100 sets of the training and the test dataset for each group. We then use the 100 sets of the training and the test dataset of each group to evaluate the average prediction accuracy of the model in the whole region (Taiwan) and individual regions.

4.3 Performance comparison

This paper uses four well-known models, such as MARS, RF, ANN, and SVR, to compare the performance with the proposed model. **Table 4** and **Table 5** report the predicting performance of the models in terms of MAE and RMSE, respectively, in different regions. The reports include individual results of 12 regions and the whole result of Taiwan. The results of the two performance measures (MAE and RMSE) are similar. Although the ANN model brings the best training performance in 12 regions, it might have overfitting problems because the test performances are worse than the other models in terms of MAE and RMSE. The performances of the MARS model, the RF model, and the SVR model are almost similar in terms of MAE

and RMSE, no matter in the training dataset and test dataset. The test performances of the proposed model are almost the best in terms of MAE and RMSE in all regions except the test MAE in Tainan. Additionally, the MAE and RMSE of the proposed model are also the best in Taiwan, and the average performance. That denotes that the proposed model brings a stable performance, whether in local or large-scale regions.

Table 4. The performance of the models in terms of MAE.

Regions	MARS		RF		ANN		SVR		The proposed model	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
Taichung	5.30	6.60	2.98	6.19	<0.01*	13.48	2.14	6.12	0.67	5.03**
Taipei	0.07	0.16	0.06	0.13	<0.01*	12.82	0.05	0.12	0.02	0.11**
Taitung	1.07	3.32	0.90	1.97	<0.01*	20.38	0.71	2.06	0.30	1.72**
Tainan	6.49	16.62	5.43	12.02	<0.01*	52.26	5.01	9.50**	3.29	9.67
Yilan	1.39	4.41	1.47	3.72	<0.01*	17.73	1.16	4.12	0.59	3.45**
Hualien	1.18	4.64	1.97	4.80	<0.01*	13.48	1.52	5.37	0.90	4.04**
Nantou	0.54	1.04	0.30	0.88	<0.01*	10.19	0.22	0.83	0.22	0.68**
Pingtung	0.80	2.40	0.92	2.47	<0.01*	13.89	0.61	2.93	0.37	1.74**
Kaohsiung	1.02	2.59	0.77	1.82	<0.01*	13.08	0.67	2.12	1.10	1.77**
New Taipei	0.06	0.11**	0.07	0.17	<0.01*	18.32	0.06	0.17	0.03	0.11**
Hsinchu	2.72	4.41	1.61	4.43	<0.01*	24.25	1.58	4.06	2.18	3.35**
Chiayi	0.21	0.30	0.13	0.31	<0.01*	11.15	0.11	0.28	0.10	0.27**
Taiwan	3.75	4.54	3.73	8.02	6.76	8.37	4.08	6.08	3.24*	3.61**
Average	1.89	3.93	1.564	3.61	0.52*	17.65	1.38	3.37	1.00	2.73**

Note: '*' denotes the best training MAE in the region; '**' denotes the best test MAE in the region.

Table 5. The performance of the models in terms of RMSE.

Regions	MARS		RF		ANN		SVR		The proposed model	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
Taichung	6.22	7.78	3.23	8.02	<0.01*	18.52	2.97	7.45	0.67	6.51**
Taipei	0.09	0.19	0.07	0.16	<0.01*	12.64	0.07	0.15	0.02	0.14**
Taitung	1.38	3.70	1.10	2.56	<0.01*	12.34	1.18	2.54	0.34	2.14**
Tainan	7.90	21.22	6.73	15.63	<0.01*	52.57	9.44	14.70	4.81	14.29**
Yilan	1.79	5.15	1.87	4.18	<0.01*	19.15	1.79	4.50	0.89	4.08**
Hualien	1.50	5.93	2.37	5.49	<0.01*	21.21	2.33	6.08	1.26	4.97**
Nantou	0.66	1.23	0.44	0.95	<0.01*	12.10	0.45	0.85	0.27	0.79**
Pingtung	0.99	2.81	1.07	2.34	<0.01*	14.24	0.82	2.82	0.40	1.94**
Kaohsiung	1.24	3.07	1.01	2.18	<0.01*	13.88	0.97	2.44	1.68	1.94**
New Taipei	0.07	0.14**	0.10	0.20	<0.01*	18.86	0.10	0.21	0.04	0.14**
Hsinchu	3.50	5.46	2.29	5.15	<0.01*	30.55	2.97	5.03	3.00	4.86**

Chiayi	0.26	0.39	0.16	0.37	<0.01*	16.77	0.18	0.33**	0.13	0.33**
Taiwan	6.56	7.89	5.09*	11.32	6.78	9.79	7.07	9.65	6.80	6.70**
Average	2.47	5.00	1.96	4.50	0.53*	19.43	2.33	4.37	1.56	3.76**

Note: '*' denotes the best training RMSE in the region; '**' denotes the best test RMSE in the region.

Fig. 5 shows the scatter plot among the actual and the predicted test dataset in Taiwan. Obviously, the prediction result of the proposed model is closest to the line. It also verifies that the proposed model is better than the other models in prediction accuracy.

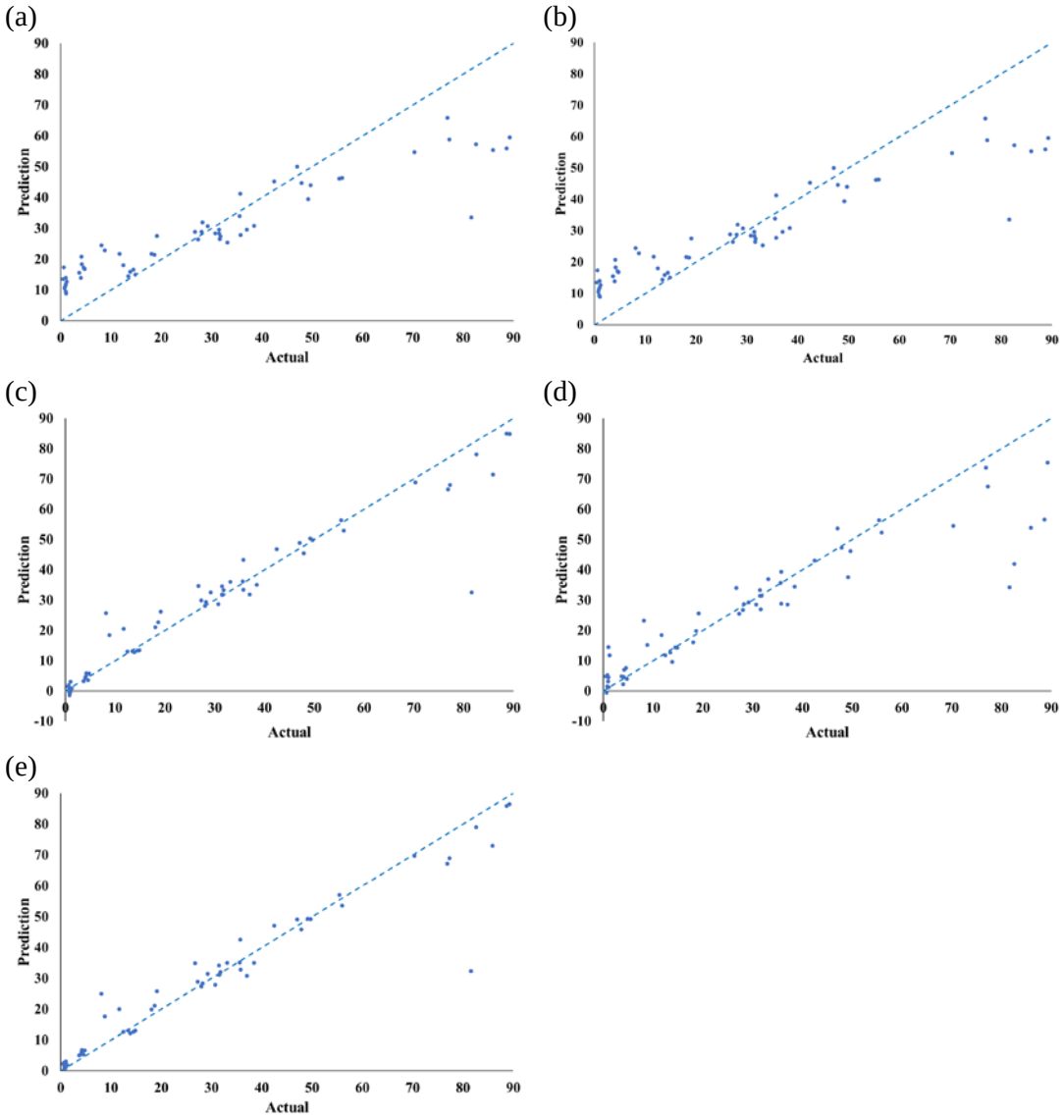

Fig. 5. The scatter plot of the actual and the predicted test dataset in Taiwan. (a) MARS; (b) RF; (c) ANN; (d) SVR; (e) The proposed model.

4.4 Discussion

Section 4.3 has reported the performance abilities of five models for predicting the rice yield in the first crop season. According to the results, we find that (a) the ANN model's test performance is the worst, even though its training performance is almost the best. Accordingly, it might have overfitting problems; (b) The performance of the MARS, RF, and SVR models are similar in terms of MAE and RMSE. The results meant that the MARS model and the RF model should be more suitable than the SVR model in predicting rice yield because the MARS model and the RF model could be used to select variables by calculating the coefficients of variables and sorting the importance of variables, respectively. Accordingly, the explanatory ability of the MARS and the RF models are both better than that of the SVR model.

On the contrary, there are two advantages to the proposed model. (a) the test performance of the proposed model is almost the best in terms of MAE and RMSE. (b) Furthermore, the test performance of the proposed model is nearly 30% better than that of the other models. This result shows that the proposed model could be applied in different regions and climatic conditions. Moreover, the proposed model brings better test performance and provides better explanatory ability because it can select the important variable for the prediction model. Accordingly, we can explain the influence of the selected variables in different regions. The results also show that the proposed model through GA selects a suitable set of climate variables to enhance the prediction performance of the SVR model.

According to the analysis results, the selected variables of the proposed model in different regions are shown in **Table 6**. The average number of selected variables is 9.8 in all regions. The cumulative rainfall should be unimportant in many regions because the rainfall variable (X_3) is less selected in many regions. On the contrary, the importance of the sunshine hours (X_1), the rainfall days (X_2), and the average temperature (X_4) should be better than the others because they are almost selected. Furthermore, the rice planting area variable for the first crop season (AREA) is selected in the whole of the rice yield regions in Taiwan. Moreover, **Table 6** also shows that the same location regions have similar selected variables, besides Kaohsiung.

Table 6. The selected variables of the proposed model.

Location	Regions	The selected variables	Variables
North Taiwan	Taipei	X_{13}, X_{14}, X_{22}, X_{23}, X_{24}, X_{32}, X_{33}, X_{42}, X_{43}, X_{44}, X_{45}, X_{52}, X_{55}, AREA	14
	New Taipei	X_{14}, X_{15}, X_{24}, X_{41}, X_{42}, X_{43}, X_{44}, X_{45}, X_{51}, X_{53}, X_{54}, AREA	12
	Hsinchu	X_{14}, X_{15}, X_{23}, X_{33}, X_{35}, X_{42}, X_{45}, X_{52}, X_{53}, AREA	10
Central Taiwan	Taichung	X_{12}, X_{13}, X_{22}, X_{23}, X_{25}, X_{32}, X_{42}, X_{45}, X_{53}, X_{55}, AREA	11
	Nantou	X_{11}, X_{13}, X_{22}, X_{23}, X_{25}, X_{32}, X_{34}, X_{43}, X_{45}, X_{53}, X_{54}, AREA	12
	Chiayi	X_{12}, X_{15}, X_{22}, X_{23}, X_{25}, X_{32}, X_{33}, X_{34}, X_{35}, X_{43}, X_{51}, X_{52}, AREA	13
South Taiwan	Tainan	X_{11}, X_{13}, X_{24}, X_{34}, X_{45}, X_{54}, X_{55}, AREA	8
	Kaohsiung	X_{44}, AREA	2
	Pingtung	X_{11}, X_{12}, X_{13}, X_{14}, X_{21}, X_{23}, X_{45}, X_{52}, X_{53}, X_{54}, AREA	11
East Taiwan	Yilan	X_{11}, X_{22}, X_{31}, X_{41}, X_{42}, X_{44}, X_{45}, X_{53}, X_{54}, AREA	10
	Hualien	X_{12}, X_{23}, X_{32}, X_{42}, X_{43}, X_{45}, X_{53}, X_{55}, AREA	9
	Taitung	X_{11}, X_{34}, X_{41}, X_{43}, X_{45}, X_{52}, X_{53}, X_{55}, AREA	9
Taiwan		X_{11}, X_{13}, X_{23}, X_{42}, X_{45}, X_{53}, AREA	7

5. Conclusions

Since the food crisis is still a serious subject and the agriculture industry plays a vital role in the stable operation of national economies, many countries must reduce the potential of food crises by making agricultural policies. Thus, many statistical prediction models have been proposed for improving prediction performance, but existing problems are limited, such as the requirement of the assumption of normality [13].

The experiment results show that the test performance of the ANN model is significantly worse than the other models in terms of all performance measures, and the ANN model might have an overfitting problem. The test performance of the proposed model in Taiwan is almost 30% better than the other models in terms of MAE and RMSE. Generally, the selected variables of the proposed model are sunshine hours, rainfall days, and the rice planting area for the first crop season. Summarized, the proposed model could bring better prediction accuracy and explain the results more clearly and directly by the selected variables. The most contribution of the proposed model is that the proposed model could be applied in different regions and according to the other climatic conditions to select suitable climatic variables to build the prediction model.

Finally, the collected data is not a large dataset in this paper due to the limitation. Accordingly, we suggest that the performance of the proposed model might be continuously improved in the future by the rice yield with more extended periods and more climate factors.

References

[1] S. Singh, "Global food crisis: magnitude, causes and policy measures," *International Journal of Social Economics*, vol. 36, no. 1/2, pp. 23-36, 2009. Article (CrossRef Link).

[2] C. P. Lee, "Reduced the Risk in Agriculture Management for Government Using Support Vector Regression with Dynamic Optimization Parameters," *Lex Localis*, vol. 15, pp. 243-261, 2017. Article (CrossRef Link).

[3] FAO, "Crop Prospects and Food Situation," 2008. Article (Web Link).

[4] FAO, "FAO Food Price Index," 2019. Article (Web Link).

[5] S. L. Ou, "Forecasting agricultural output with an improved grey forecasting model based on the genetic algorithm," *Computers and Electronics in Agriculture*, vol. 85, pp. 33-39, 2012. Article (CrossRef Link).

[6] T. Xiong, C. Li, Y. Bao, Z. Hu, and L. Zhang, "A combination method for interval forecasting of agricultural commodity futures prices," *Knowledge-Based Systems*, vol. 77, pp. 92-102, 2015. Article (CrossRef Link).

[7] I. Yunita, G. Taib, and R. A. Hadiguna, "Coffee bean supply chain strategy: the case of trading institution and profit margin for pioneer coffee commodities in Indonesia," *International Journal of Agriculture Innovation, Technology and Globalisation*, vol. 1, no. 1, pp. 57-66, 2019. Article (CrossRef Link).

[8] T. Y. Chang, "The influence of agricultural policies on agriculture structure adjustment in Taiwan: An analysis of off-farm labor movement," *China Agricultural Economic Review*, vol. 3, no. 1, pp. 67-79, 2011. Article (CrossRef Link).

[9] R. Mano and C. Nhemachena, "Assessment of the Economic Impacts of Climate Change on Agriculture in Zimbabwe: A Ricardian Approach," *The World Bank, Policy Research Working Paper Series*, 2007. Article (CrossRef Link).

[10] F. J. Meza, J. W. Hansen, and D. Osgood, "Economic Value of Seasonal Climate Forecasts for Agriculture: Review of Ex-Ante Assessments and Recommendations for Future Research," *Journal of Applied Meteorology and Climatology*, vol. 47, no. 5, pp. 1269-1286, 2008. Article (CrossRef Link).

[11] S. M. Say, M. Keskin, M. Sehri, and Y. E. Sekerli, "Adoption of precision agriculture technologies in developed and developing countries," *The Online Journal of Science and Technology*, vol. 8, no. 1, pp. 7-15, 2018. Article (CrossRef Link).
[12] T. Jheng, T. Li, and C. Lee, "Using hybrid support vector regression to predict agricultural output," in *Proc. of 27th Wireless and Optical Communication Conference (WOCC)*, pp. 1-3, 2018. Article (CrossRef Link).
[13] C. P. Lee, G. J. Shieh, T. J. Yiu, and B. J. Kuo, "The strategy to simulate the cross-pollination rate for the co-existence of genetically modified (GM) and non-GM crops by using FPSOSVR," *Chemometrics and Intelligent Laboratory Systems*, vol. 122, pp. 50-57, 2013. ∈ Article (CrossRef Link).
[14] Y. Chung, D. Choi, H. Choi, D. Park, H.-H. Chang, and S. Kim, "Automated Detection of Cattle Mounting using Side-View Camera," *KSII Transactions on Internet and Information Systems*, vol. 9, no. 8, pp. 3151-3168, 2015. Article (CrossRef Link).
[15] N. M. Ranka and L. Sharma, "Design of experiments: a powerful tool for agriculture analysis," *Elixir Statistics*, vol. 52, pp. 11356-11358, 2012. Article (CrossRef Link).
[16] D. J. Street, "Fisher's Contributions to Agricultural Statistics," *Biometrics*, vol. 46, no. 4, pp. 937-945, 1990. Article (CrossRef Link).
[17] M. Kaul, R. L. Hill, and C. Walthall, "Artificial neural networks for corn and soybean yield prediction," *Agricultural Systems*, vol. 85, no. 1, pp. 1-18, 2005. Article (CrossRef Link).
[18] J. Y. Shin, K. R. Kim, and J. C. Ha, "Seasonal forecasting of daily mean air temperatures using a coupled global climate model and machine learning algorithm for field-scale agricultural management," *Agricultural and Forest Meteorology*, vol. 281, p. 107858, 2020. Article (CrossRef Link).
[19] Y. Tian, Y. P. Xu, and G. Wang, "Agricultural drought prediction using climate indices based on Support Vector Regression in Xiangjiang River basin," *Science of The Total Environment*, vol. 622-623, pp. 710-720, 2018. Article (CrossRef Link).
[20] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995. Article (CrossRef Link).
[21] A. Jaafari and H. R. Pourghasemi, "28 - Factors Influencing Regional-Scale Wildfire Probability in Iran: An Application of Random Forest and Support Vector Machine," in *Spatial Modeling in GIS and R for Earth and Environmental Sciences*, H. R. Pourghasemi and C. Gokceoglu Eds.: Elsevier, 2019, pp. 607-619. Article (CrossRef Link).
[22] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support Vector Regression Machines," *the Neural Information Processing Systems (NIPS)*, pp. 155-161, 1996. Article (CrossRef Link).
[23] D. Basak, S. Pal, and D. C. Patranabis, "Support Vector Regression," 2007. Article (CrossRef Link).
[24] C.-P. Lee, W.-C. Lin, and C.-C. Yang, "A strategy for forecasting option price using fuzzy time series and least square support vector regression with a bootstrap model," *Scientia Iranica*, vol. 21, no. 3, pp. 815-825, 2014. Article (CrossRef Link).
[25] C. H. Huang, F. H. Yang, and C. P. Lee, "The strategy of investment in the stock market using modified support vector regression model," *Scientia Iranica*, vol. 25, no. 3, pp. 1629-1640, 2018. Article (CrossRef Link).
[26] J. H. Holland, *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*, MIT Press, 1992. Article (CrossRef Link).
[27] S. Mirjalili, "Genetic Algorithm," in *Evolutionary Algorithms and Neural Networks: Theory and Applications*, Cham: Springer International Publishing, pp. 43-55, 2019. Article (CrossRef Link).
[28] B. Babar, L. T. Luppino, T. Boström, and S. N. Anfinsen, "Random forest regression for improved mapping of solar irradiance at high latitudes," *Solar Energy*, vol. 198, pp. 81-92, 2020. Article (CrossRef Link).

[29] K. Millard and M. Richardson, "On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping," *Remote Sensing*, vol. 7, no. 7, pp. 8489-8515, 2015. Article (CrossRef Link).
[30] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001. Article (CrossRef Link).
[31] E. Izquierdo-Verdiguier and R. Zurita-Milla, "An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 88, p. 102051, 2020. Article (CrossRef Link).
[32] J. C. W. Chan and D. Paelinckx, "Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery," *Remote Sensing of Environment*, vol. 112, no. 6, pp. 2999-3011, 2008. Article (CrossRef Link).
[33] D. C. Lo, C. C. Wei, and E. P. Tsai, "Parameter Automatic Calibration Approach for Neural-Network-Based Cyclonic Precipitation Forecast Models," *Water*, vol. 7, no. 7, pp. 3963-3977, 2015. Article (CrossRef Link).
[34] G. M. Mulualem and Y.-A. Liou, "Application of Artificial Neural Networks in Forecasting a Standardized Precipitation Evapotranspiration Index for the Upper Blue Nile Basin," *Water*, vol. 12, no. 3, p. 643, 2020. Article (CrossRef Link).
[35] C. C. Yang, Y. Leu, and C. P. Lee, "A Dynamic Weighted Distancedbased Fuzzy Time Series Neural Network with Bootstrap Model for Option Price Forecasting," *Journal of Economic Forecasting*, no. 2, pp. 115-129, 2014. Article (CrossRef Link).
[36] J. H. Friedman, "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1-67, 1991. Article (CrossRef Link).
[37] W. Zhang and A. T. C. Goh, "Multivariate adaptive regression splines and neural network models for prediction of pile drivability," *Geoscience Frontiers*, vol. 7, no. 1, pp. 45-52, 2016. Article (CrossRef Link).
[38] S. Xu et al., "Spatial Downscaling of Land Surface Temperature Based on a Multi-Factor Geographically Weighted Machine Learning Model," *Remote Sensing*, vol. 13, no. 6, p. 1186, 2021. Article (CrossRef Link).
[39] A. Gani, K. Mohammadi, S. Shamshirband, T. A. Altameem, D. Petković, and S. Ch, "A combined method to estimate wind speed distribution based on integrating the support vector machine with firefly algorithm," *Environmental Progress & Sustainable Energy*, vol. 35, no. 3, pp. 867-875, 2016. Article (CrossRef Link).
[40] C. P. Lee and Y. Leu, "A novel hybrid feature selection method for microarray data analysis," *Applied Soft Computing*, vol. 11, no. 1, pp. 208-213, 2011. Article (CrossRef Link).
[41] S. H. Wu, "Forecasting of Typhoon Path Raided in Taiwan by Using Neural Network Models," *Department of Civil Engineering, National Pingtung University of Science and Technology*, 2014. Article (CrossRef Link).

Chin-Hung Kuan is currently a Ph.D. candidate in the Department of Information Management, National Taiwan University of Science and Technology, Taipei City, Taiwan. His research interests include Artificial Intelligence, Big data analysis.

Yungho Leu received his B.S. and M.S. degrees in Electrical Engineering in 1981 and 1983, respectively, from National Taiwan University, Taiwan. He received his Ph.D. degree in Computer Science in 1991 from Purdue University, USA. He is now a professor in the Department of Information Management, National Taiwan University of Science and Technology. His current research interests include quality control in semiconductor manufacturing, Deep Learning, Machine Learning, Natural Language Processing, and parallel frequent itemset mining. He is a senior member of IEEE.

Chien-Pang Lee received his B.S. degree in Applied Statistics in 2003 from Ming Chuan University, Taiwan, his M.S. degree in Biostatistics in 2006 from Nation Chung Hsing University, Taiwan, and his Ph.D. degree in Information Management in 2010 from Nation Taiwan University of Science and Technology, Taiwan. He is currently an Associate Professor in the Department of Maritime Information and Technology, National Kaohsiung University of Science and Technology. His research interests include Big data analysis, Bioinformatics, Statistics, and Financial forecasting.