Agriculture and Food

AIMS Agriculture and Food, 8(4): 1052–1070.
DOI: 10.3934/agrfood.2023057
Received: 23 June 2023
Revised: 26 September 2023
Accepted: 01 November 2023
Published: 11 November 2023

http://www.aimspress.com/journal/agriculture

Research article

Forecasting arabica coffee yields by auto-regressive integrated moving average and machine learning approaches

Yotsaphat Kittichotsatsawat[1], Anuwat Boonprasope[1], Erwin Rauch[2], Nakorn Tippayawong[3] and Korrakot Yaibuathet Tippayawong[1,4,*]

[1] Supply Chain and Engineering Management Research Unit, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Industrial Engineering, Free University of Bolzano, Bolzano, Italy
[3] Department of Mechanical Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
[4] Department of Industrial Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand

* **Correspondence:** Email: korrakot@eng.cmu.ac.th; Tel.: +66816719019.

Abstract: Coffee is a major industrial crop that creates high economic value in Thailand and other countries worldwide. A lack of certainty in forecasting coffee production could lead to serious operation problems for business. Applying machine learning (ML) to coffee production is crucial since it can help in productivity prediction and increase prediction accuracy rate in response to customer demands. An ML technique of artificial neural network (ANN) model, and a statistical technique of autoregressive integrated moving average (ARIMA) model were adopted in this study to forecast arabica coffee yields. Six variable datasets were collected from 2004 to 2018, including cultivated areas, productivity zone, rainfalls, relative humidity and minimum and maximum temperatures, totaling 180 time-series data points. Their prediction performances were evaluated in terms of correlation coefficient (R^2), and root means square error (RMSE). From this work, the ARIMA model was optimized using the fitting model of (p, d, q) amounted to 64 conditions through the Akaike information criteria arriving at (2,1,2). The ARIMA results showed that its R^2 and RMSE were 0.7041 and 0.1348, respectively. Moreover, the R^2 and RMSE of the ANN model were 0.9299 and 0.0642 by the Levenberg-Marquardt algorithm with TrainLM and LearnGDM training functions, two hidden layers and six processing elements. Both models were acceptable in forecasting the annual arabica coffee production, but the ANN model appeared to perform better.

Keywords: AI; agricultural planning; digital agriculture; SME4.0; time series analysis

1. Introduction

In the agricultural sector, there remain challenges in agricultural management in response to customer needs. The cause of this problem is from lack of know-how and knowledge management [1–3]. These issues indicate that entrepreneurs need effective tools for developing and increasing productivity in the production process for long-term stability. Additionally, inaccurate yield prediction can have far-reaching effects on food production, supply systems, economies and global food security. Uncertainty in predicting crop yields and quality can lead to lower crop yields, reduced income, financial instability for agricultural producers, increased production costs, shortages and price fluctuations [4]. In marketing, it affects price volatility, suboptimal policy choices and resource misallocation and disrupts international trade agreements and negotiations, affecting trade balances and economic stability [5]. Accurate forecasting is more critical for farmers who need to adapt their practices to changing climate conditions [6].

We focused on arabica coffee (*Coffea arabica L.*) grown in northern Thailand. It is among the most popular species of coffee due to its features and flavor that offer superior quality than other types [7]. Currently, the coffee business is becoming more competitive. Its production forecast is of great interest to stakeholders involved. A lack of certainty in forecasting coffee production is especially vulnerable, affecting the whole supply chain from coffee farmers to exporters, importers, roasters and retailers, leading to supply gaps, disappointing customers and potentially damaging brand reputation [8]. Furthermore, price volatility will affect their profitability and financial planning [9]. Uncertain forecasts can lead to overstocking or understocking [10]. Coffee cultivation is usually long-term. If the farmer or entrepreneurs lack knowledge of the processing management, it will result in the uncertainty in production forecasts [11].

To address these challenges in coffee businesses, artificial intelligence (AI) is considered essential for modern manufacturing processes in agriculture and industry. This technology is likely to generate increased efficacy and effectiveness in the production process to enhance companies' potential according to international standards. Over the past few years, the agriculture and industry sectors have introduced various technologies to increasingly modernize their manufacturing and agricultural operations. AI is also being used to analyze the data and accelerate the operating system with flexibility that leads to more effectiveness in producing products or services in accordance with customer needs [12–14].

The artificial neural network (ANN), an algorithm of machine learning (ML) model, is viewed as a vital data-modeling tool [15]. It can calculate the data through the functional structures of neural networks. Input and output data processes are run through a neuron network, including single-layer perceptions, multilayer perceptions, recurrent ANN and self-organization mapping [16]. Recently, many agriculture sectors applied the ANN to predict the productivity of products [17,18]. Kittichotsatsawat et al. [19] used basic ANN models to predict the productivity of coffee in northern Thailand. Bhojani et al. [20] applied ANN to wheat yield prediction. Palanivel et al. [21] also utilized ANN to predict crop yield. Important factors considered include the area, productivity zone, rainfall, relative humidity, temperature, etc. [22–26]. It is noted here that, apart from applying ML models, other methods can help to improve the productivity of cherry coffee. Examples are the

analytical hierarchy process (AHP) and frequency ratio (FR), attention mechanism (AM), convolutional neural network (CNN), hyperspectral image (HSI), spectral–spatial features from principal component analysis (PCA), weighted linear combination (WLC), attitude determination and control subsystem (ADCS) models [27–30].

Autoregressive integrated moving average (ARIMA) model is one of the statistical tools that has been used to predict various agricultural product output. It can detect the data through Box-Jenkins in order to create the ARIMA model, including (i) stationary, (ii) co–integration and (iii) error correction mechanism [31]. Padhan [32] employed ARIMA to forecast agricultural productivity in India. ARIMA was used to predict the productivity of goods in terms of area, zone, relative humidity, rainfall, temperature, etc. [33–36].

Techniques such as ANN and ARIMA may be used to determine the productivity of coffee to meet customer requirements. From the literature review, there have been several works applying the ANN and ARIMA models to predict agricultural productivity, such as commodities, agricultural products, crop price, etc. [37–39]. However, it was noticed that ARIMA and ANN are not yet employed in coffee production forecast. Therefore, we aim to predict the cherry coffee yield and compare the performance between the ARIMA and ANN models. This finding will benefit and help in analyzing the trends of Thai coffee effectively and sustainably.

2. Materials and methods

2.1. Data collection and treatment

In this study, datasets were collected for 15 years from 2004 to 2018 (180 months). The input predictor data considered were from the Thai Agricultural Economics Office and the Meteorological Department. They included the cultivated area, productivity zone, monthly rainfall, monthly RH and monthly temperatures. Furthermore, the output data was from the productivity yield each year. It was noted that the coffee was yielded only six months in a year with increasing trend for the past several years, shown in Figure 1.

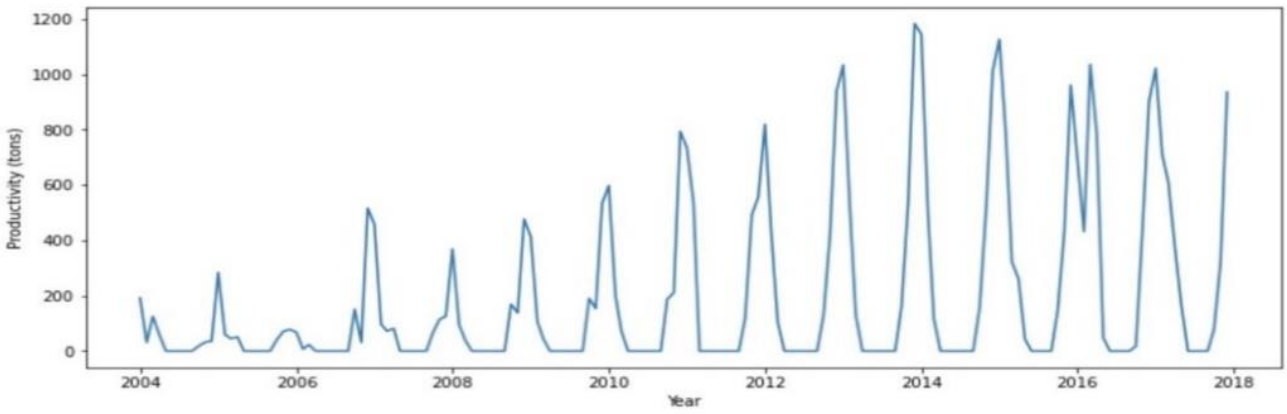

Figure 1. Historical data of arabica coffee production in Thailand.

Based on the literature, it is necessary to normalize and standardize the values of input features

and output targets before developing ML models [40,41]. In this work, the input and output variables are normalized in the range 0–1, using:

$$N = \frac{(X - X_{min})}{(X_{max} - X_{min})} \quad (1)$$

where N is the normalized data; X is the measured value: X_{min} and X_{max} are the minimum and maximum values.

The ARIMA and ANN performances had to be measured according to a validation of variables dataset. The ANN selected was tested. Then, the coefficient of determination (R^2), the root means square error (RMSE) [42,43] as well as the mean squared error (MSE), were compared.

$$SST = \sum_{i=1}^{n}(y - \bar{y})^2 = (y - \overline{y})' (y - 1\bar{y}) \quad (2)$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = (\hat{y} - \overline{y})' (\hat{y} - 1\bar{y}) \quad (3)$$

$$R^2 = 1 - \frac{SS_{regress}}{SS_{total}} \quad (4)$$

where y_i and $\hat{y}$ are the square of the sample correlation, $SS_{regression}$ is the sum of squares due to regression (explained sum of squares) and SS_{total} is the total sum of squares.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[E(x_i) - M(x_i)]^2} \quad (5)$$

$$\text{MSE} = (\text{RMSE})^2 \quad (6)$$

where n is the sample size of the testing dataset, while $E(x_i)$ and $M(x_i)$ are interpolated/predicted and observed values, respectively.

2.2. Forecasting techniques

A prediction is built on the foundation of some scientific calculation based on historical data. The variable datasets were analyzed through the ARIMA model with Python programming and ANN model using MATLAB programming.

2.2.1. ARIMA model

ARIMA model is a technique of statistics and econometrics that evaluates the events that will happen over each period of time.

$$y_t = \theta_0 + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}, \quad (7)$$

where y_t and ε_t are the actual value and random error at time period t, respectively; ϕ_i (I = 1, 2…., p) and θ_j (j = 0,1, 2…, q) are model parameter. p and q are integers and often referred to as order of the model. Random errors, ε_t, are assumed to be independently and identically distributed with a mean of zero and a constant variance of σ^2.

$$Y_t = f(t) + \varepsilon_t \quad (8)$$

where, Y_t signifies production for the time t in year, f (t) denotes a function of time t and ε_t denotes

production error (i.e., the difference between observed and forecasted production for time *t* year). Once a functional link between production and time (in other words, a time series model) has been built, production for year t + 1 can be forecasted. The first stage in creating this model is determining whether the time series under consideration is stationary or non-stationary.

$$\varphi_{p(B)}\Delta^d h_t = c + \theta_q(B) g_t \tag{9}$$

where, h_t is variable under forecasting at time *t*, B is lag operator, g is error term (Y-$\hat{Y}$ in which $\hat{Y}$ is the estimated value of Y), $\varphi_p(B)$ is non-seasonal AR i.e., the autoregressive operator, represented as a polynomial in the back shift operator, (1 − B) *d* is non-seasonal difference, θ_q(B) is non-seasonal moving average i.e., the moving average operator, represented as a polynomial in the backshift operator, φ′ s and θ′ s are the parameters to be estimated.

The variable datasets were prepared to consider the time series component, including trend, season, cycle and irregularity. ARIMA model was split into two parts, with 156 data for training and 24 for testing. The historical observations and random mistakes (errors) were used to estimate the future variables dataset. It was shown on Box-Jenkins to predict the future value through ARIMA modeling, including a three-step iterative technique (i) model identification, (ii) parameter estimation and (iii) residual diagnostics testing [44].

The unit root test series graphs revealed the autocorrelation function (ACF) and partial ACF (PACF) [45]. The zig-zag trend will show the increase to meet the stationary series graphs. After the stationary time series process was identified, the ARIMA model was defined by the autoregressive integrated moving average model (p, d, q). Python programming was used to detect a suitable residual of the ACF graph with a 95% confidence band [46]. Next, a suitable ARIMA model was used in prediction (156 data for training and 24 data for testing).

2.2.2. ANN model

ANN is a complex multivariate model to approximate the unknown expectation function of a random variable. Weights will be used to estimate the parameters in the ANN model.

$$t_i = \sum_{j-1}^{n} W_{ij} X_j \tag{10}$$

where *n* is number of inputs, *w* is weight of the connection between i^{th} and j^{th} node and *x* is input from node *j*. Calculation of output will be analyzed through a transfer function of O_i;

$$O_i = f(t_i) \tag{11}$$

The variable datasets are randomly separated into three groups to prevent overfitting. The variable data were divided randomly whose 70% used for training, 15% for validating and 15% for testing [47]. During ANN training, the algorithm was furnished with the performance of minimum or maximum through the shortest path in order to gain the network's yield size. The neural network performance was accomplished by backpropagation via the training set in order to update the minimum MSE during the training set [48].

The neural networks were trained by means of a training set, and the output datasets were compared with the fixed weight. Then, feed-forward backpropagation was utilized to compare output

datasets with the fixed weight. The MSE was utilized to test and calculate epochs to validate the variable dataset in the part of the neural network running. Neurons of the neural network will utilize a definite function in the hidden layer and gather the combination and bias. Lastly, the output of variable data will give the predicted model [49]. The crop yield index was determined via the input and output variables set. During the neural network process, each independent variable set was assessed and revealed by a partial dependence plot (PDP) [50,51]. The highest importance value showed the relative importance of that parameter from each index.

2.2.3. Prediction performance evaluation

The crop yield of cherry coffee was validated through the input variables dataset based on the difference between observed and predicted coffee crops. The leave-one-out cross-validation technique was used to be randomly evaluated with the ANN model, while the ARIMA model used time-based cross-validation. The 156 months were evenly partitioned in order to evaluate the cross-validation. Three rounds of ANN and ARIMA were randomly evaluated throughout variable datasets. The variable datasets were trained and treated individually in each dataset. Finally, the performance model was evaluated and showed the RMSE and MSE. The best model was determined through the largest R^2 and the smallest RMSE.

3. Results

3.1. ARIMA prediction results

3.1.1. Identifying the data type

Time series analysis of ARIMA was completed based on the 180 monthly input variables dataset. Variable datasets were detected through the unit root with stationary test in order to examine the stationary or non-stationary data. A statistical test was used to consider these data, specifically the Augmented Dickey-Fuller (ADF) test, and based on the p-value of the ADF-test, if the result shows a value less than 0.05, it will be identified as stationary [52].

After the stationary test, the p-value of 0.9210 was obtained and variable datasets were non-stationary. However, when examining the data with ACF and PACF as in Figure 2, it was shown that the time series was revealed as seasonal and stationary. Nonetheless, the data was adjusted and rearranged to investigate the difference of the first number (d = 1). This time, the p-value was 1.426 x 10^{-13}, thus, the variable data was identified as stationary. However, the ACF and PACF showed the space of seasonal components (12, 24, 36 units) in Figure 3.

3.1.2. Optimizing ARIMA model parameters

The data was divided into two parts, including 156 for training data and 24 for testing data. Then, tuning the model by training data through ARIMA model fitting was carried out in order to define the parameters through 64 conditions based on the autoregressive (AR(p)), integrated (I(d)) and moving average (MA(q)) [53]. The Akaike information criteria (AIC) was used to return the conditions of each value [54], as shown in Table 1.

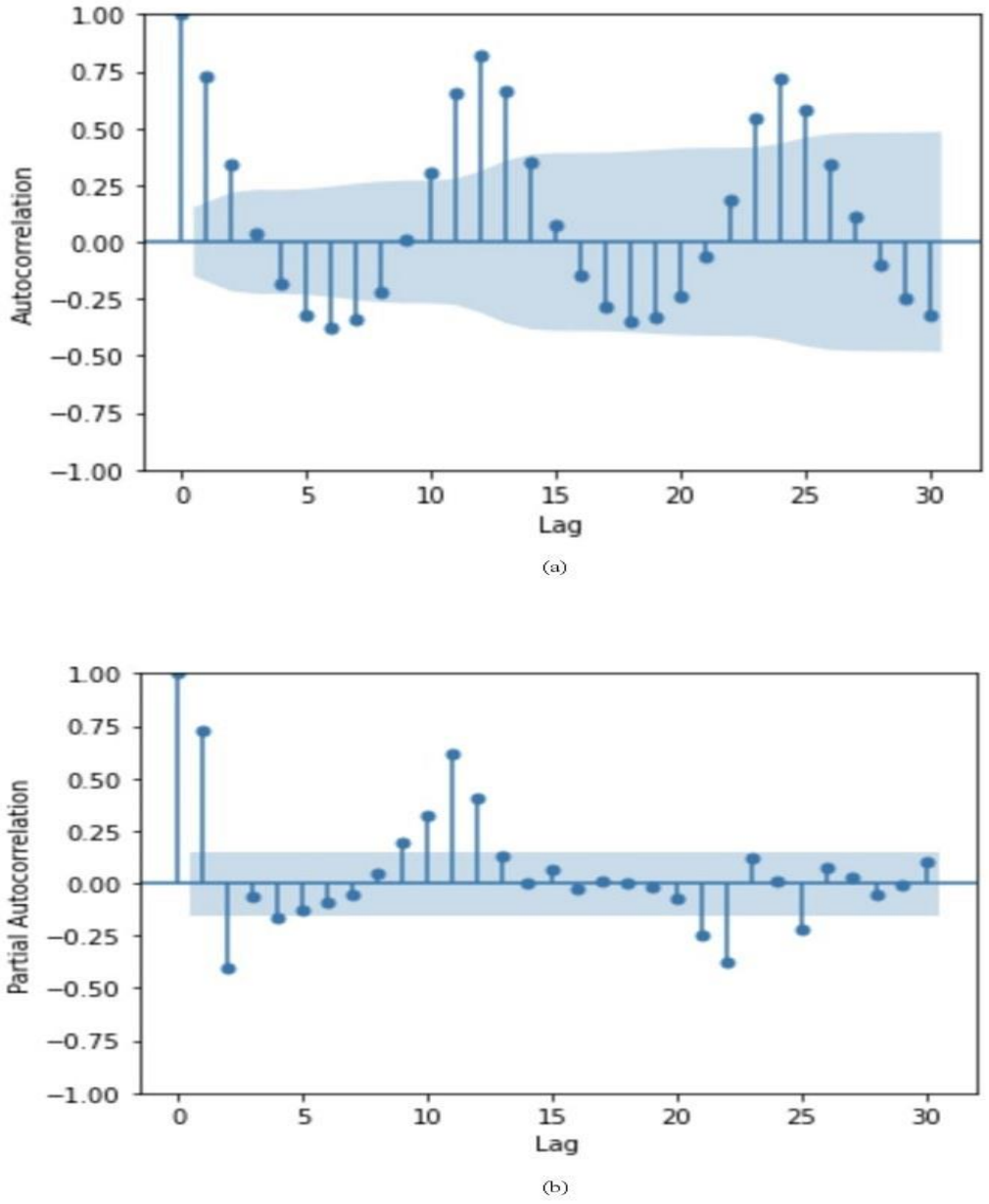

Figure 2. Autocorrelation function (ACF) and partial ACF of coffee prediction.

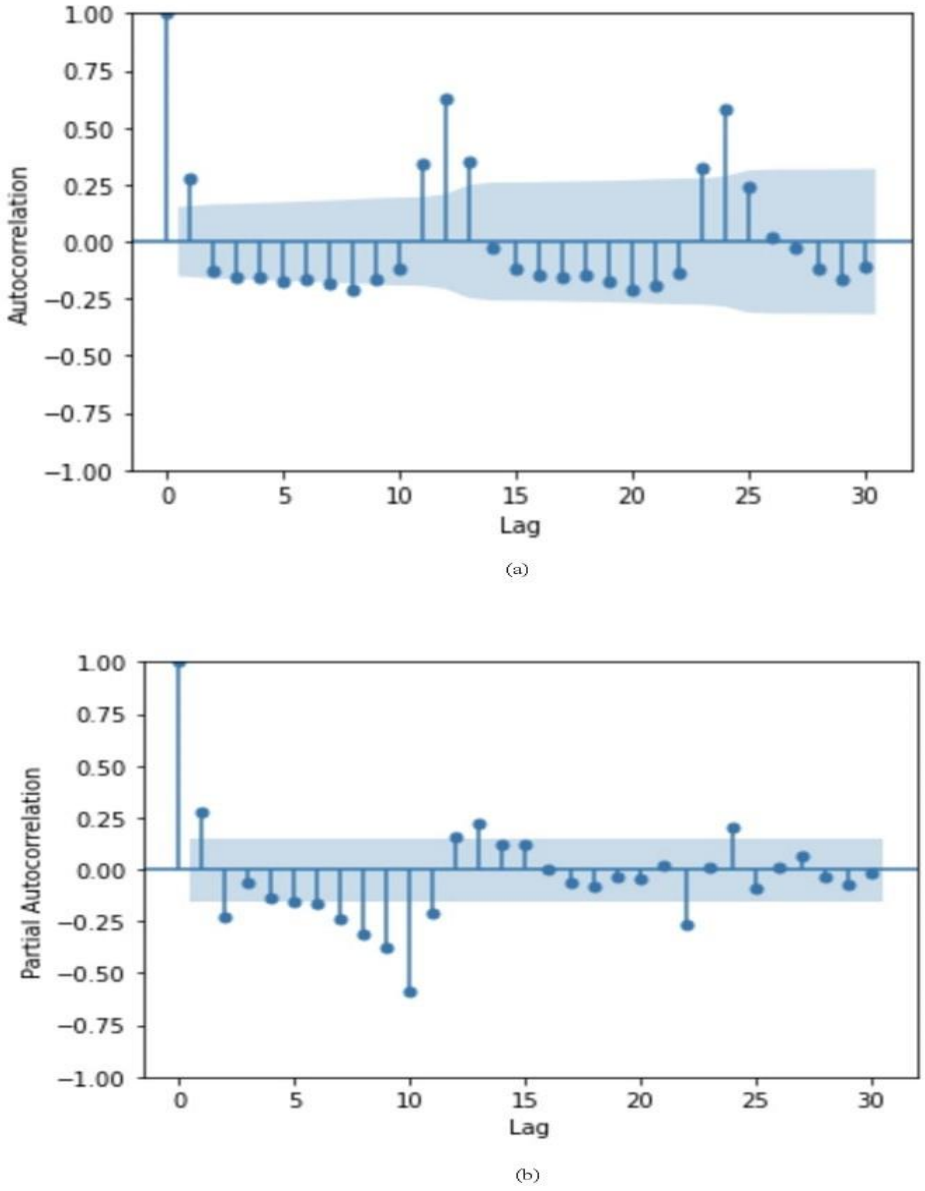

Figure 3. Autocorrelation function (ACF) and partial ACF of coffee prediction through adjusting and rearranging data.

Table 1. Optimized ARIMA model parameters of coffee yields.

No.	Parameter	AIC
1	(2,1,2)	−168.0802
2	(2,1,3)	−166.9035
3	(2,0,3)	−166.7370
4	(3,1,2)	−166.4853
5	(3,0,2)	−165.7990

Table 1 shows the five most minor (p, d, q) condition parameters; the AIC of (p, d, q) is 168.0802 at (2,1,2). However, when (p, d, q) parameters were identified, it led to cross validation based on time-based cross-validation. The data was changed from random sampling to one by one through a training model or forward chaining by identifying the data ratio between training data and data of testing amounted 12 rounds.

In Table 2, the ARIMA model shows R^2 of 0.0741, RMSE of 0.1348, and MSE of 0.0181. However, the relation of target and output of variable datasets, which is the trends of the data association, was unidirectional, as shown in Figure 4. While the results of cross-validation showed an average R^2 to be higher, while the average of MSE of test data was smaller. So, this model configuration is suitable, and it can be used in prediction.

Table 2. R^2 and MSE of training and testing and time-based cross-validation of coffee yields.

Training and Testing						Time-based cross-validation			
MSE Train	MSE Test	RMSE Train	RMSE Test	R^2 Train	R^2 Test	R^2 Train	R^2 Test	MSE Train	MSE Test
0.0181	0.0469	0.1348	0.2167	0.7041	0.3521	0.7383	0.3931	0.0046	0.0451

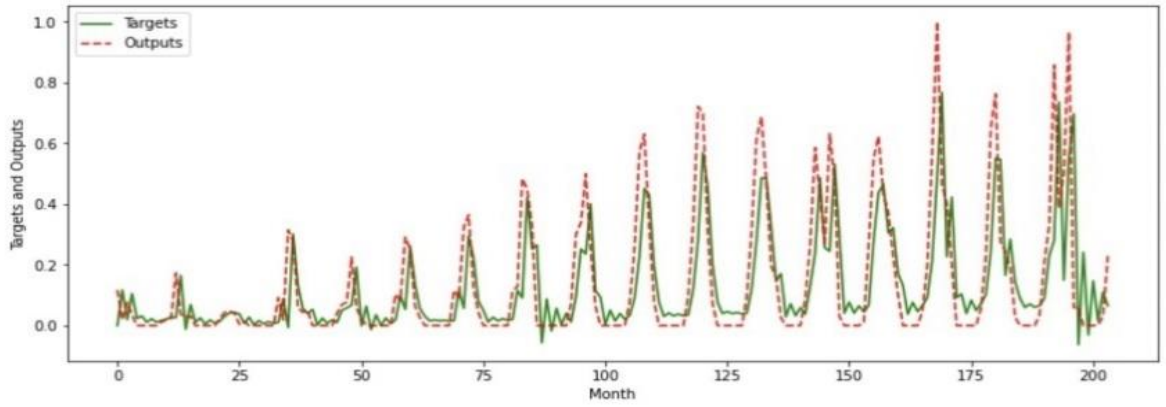

Figure 4. Targets and outputs for the coffee prediction.

3.2. ANN prediction results

The ANN analysis was achieved based on the 180 monthly input data. The variable datasets included the cultivated area (X_1) for the coffee. The productivity zone (X_2) is the factor that implies the quantity of cherry coffee in each crop. The rainfall data (X_3) is a crucial factor for the output of coffee in each year. The proper amount of RH (X_4) will enable the high quantity of coffee yield. The maximum and minimum ambient temperatures (X_5 and X_6) are essential to the coffee productivity.

For ANN hyperparameter setting, the number of hidden layers and the number of neurons (i.e., processing elements (PEs)) for each hidden layer were optimized by trial and error in predicting the coffee yields. The MSE, RMSE and R^2 values were used to evaluate the optimal parameters. From Table 3, the network properties were arranged with (a) network type of feed-forward backpropagation, (b) training and adaptation learning functions through Levenberg-Marquardt algorithm with TRAINLM and LEARNGDM, (c) MSE of performance function, (d) varying hidden layers and six

neurons and (e) TANSIG transfer function [55]. After that, the performances of the ANN models to predict cherry coffee productivity yield were evaluated for various ANN configurations. The best training results of the ANN model were two hidden layers and one PE for each hidden layer, which provided the R values of the training, testing, validating data phases to be 0.9921, 0.9384 and 0.8723, respectively, and the MSE of the validating data to be 19576, as also shown in Figures 5 and 6.

Figure 5 shows the validation performance with the MSE value of the validating data for the best training results of the ANN model. The learning rate was set to be 0.02, while the learning cycles of the model was done to be 1000 epochs. The optimal validation performance of the ANN model with two hidden layers and one PE each was found at 13 epochs, which was based on the lowest MSE value of the validating data for this ANN configuration.

Figure 6 shows the predicted values versus measured values for training, testing, validating data and whole data. The perfect prediction established the accuracy of the neural network in predicting the cherry coffee productivity yield based on the calculation of the index of the variable.

Table 3. Performances of various ANN configurations through MSE, R training, R testing, R validation, R overall and R^2.

Number of hidden layers	PEs	MSE	R Training	R Testing	R Validation	R Overall	R^2
1	1	20791	0.9014	0.5100	0.9196	0.8491	0.7210
1	2	27138	0.7977	0.7603	0.5115	0.7764	0.6028
1	3	40226	0.8079	0.8405	0.7176	0.7965	0.6344
1	4	42773	0.7506	0.3192	0.7525	0.7096	0.5035
1	5	50880	0.8474	0.7186	0.4740	0.7795	0.6076
1	6	39929	0.8580	0.6132	0.7423	0.7794	0.6075
1	7	72995	0.9761	0.6014	0.4396	0.8464	0.7164
1	8	20092	0.7720	0.6825	0.6804	0.7431	0.5522
1	9	20419	0.9798	0.8299	0.8797	0.9291	0.8632
1	10	15591	0.9810	0.5857	0.9452	0.9000	0.8100
2	1	19576	0.9921	0.9384	0.8723	0.9643	0.9299
2	2	20384	0.7900	0.7736	0.6675	0.7807	0.6095
2	3	14133	0.7776	0.8219	0.7922	0.7747	0.6002
2	4	27636	0.9143	0.5913	0.8700	0.8816	0.7772
2	5	28923	0.9297	0.7595	0.9057	0.9085	0.8254
2	6	16209	0.9131	0.7751	0.7987	0.8690	0.7552
2	7	37072	0.7236	0.6636	0.6070	0.7042	0.4959
2	8	16657	0.9825	0.9049	0.8930	0.9629	0.9272
2	9	7218	0.8151	0.6677	0.8071	0.8039	0.6463
2	10	16288	0.9761	0.6510	0.8018	0.9164	0.8398

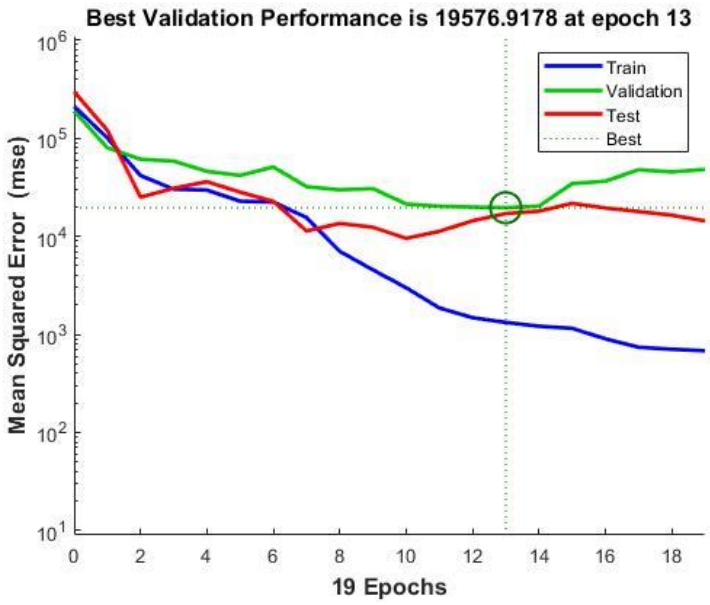

Figure 5. Validation performance with the MSE value of the validating data.

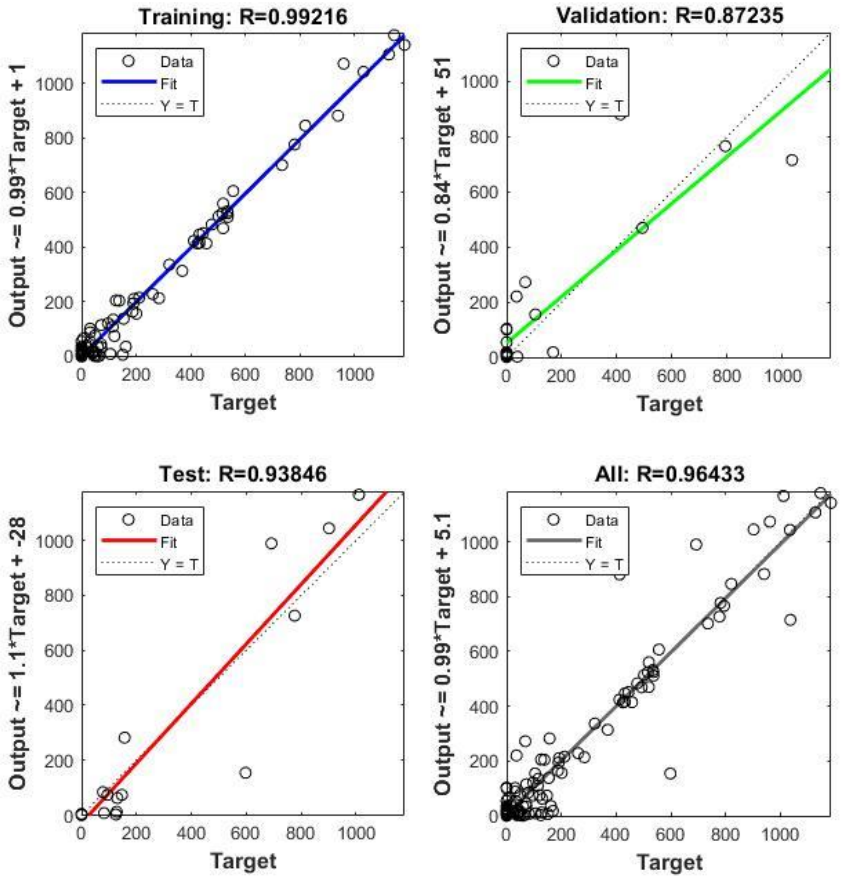

Figure 6. ANN modeling in coffee yield (ton) prediction.

Figure 7 shows one-way partial dependence plots (PDPs) for each variable's relative importance. The monthly temperature (X_5) (Figure 7e) showed a higher effect on a variable dataset in this model, with the PDP value varied from 0.0501 to 0.5211. The second most important predictor was the productivity zone (X_2) (Figure 7b), with the PDP value from 0.0896 to 0.2929. Similarly, the third crucial variable was revealed to be monthly rainfall (X_3) (Figure 7c), showing the PDP value from 0.0899 to 0.2760. Moreover, the cultivated area (X_1) (Figure 7a) and minimum monthly temperature (X_6) (Figure 7f) showed the PDP value of small difference, which were 0.1373 to 0.2010 and 0.1883 to 0.1275, respectively. Lastly, the relative humidity (X_4) showed marginal effects on the model PDP with values from 0.1295 to 0.1670.

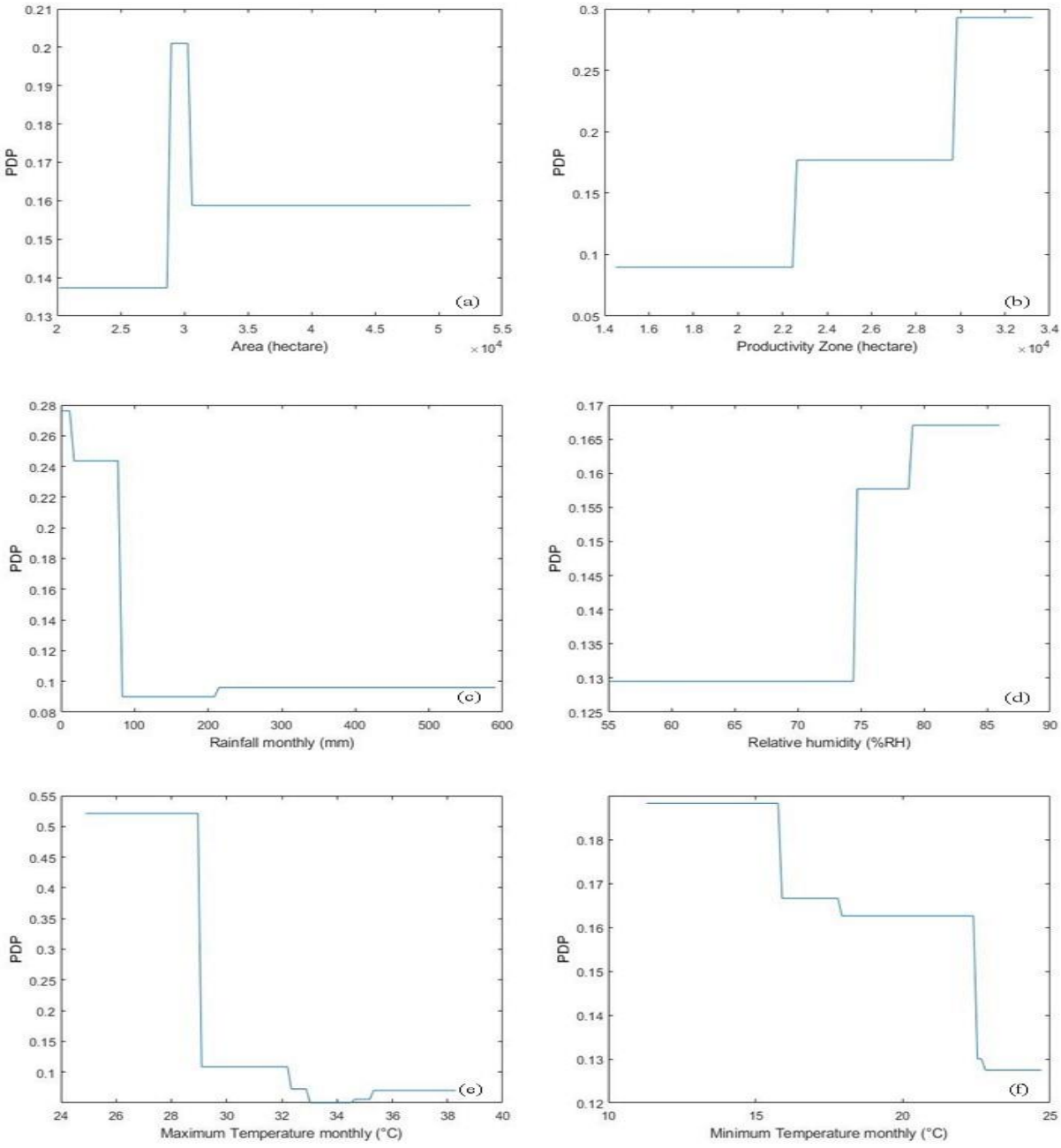

Figure 7. One-way PDPs of coffee yield prediction.

The maximum temperature (e) affected the amount of productivity significantly. If the temperature was higher than or equal to 29 ℃, the productivity was decreased. If the minimum temperature (f) was less than or equal to 15–20 ℃, the productivity was improved. The rainfall (c) was one of the essential factors because the coffee productivity depended on the amount of rainfall each year. A suitable rainfall should be less than 100 mm, leading to good coffee plantation condition. The productivity zone (b) and cultivated area (a) were directly affected the quantity of coffee production If the farmers have more productivity zone and area, they will have higher production. Finally, relative humidity (d) should be high because it is preferable for coffee cultivation.

4. Discussion

Yield of arabica coffee is relatively unstable due to many factors, for example, changing weather conditions, different soil pH, fluctuation of ambient temperature, alteration of moisture in air, etc. Therefore, it is essential to forecast the coffee productivity to go along with customer' expectations.

In this study, ARIMA and ANN were deployed to analyze and predict the crop yield of arabica coffee using data from 2004 to 2018. Both models have been demonstrated to be efficient in forecasting coffee production. The prediction performances of these models were evaluated using R^2 and RMSE. The ARIMA model was optimized for (p, d, q) at (2,1,2). Its R^2 and RMSE were 0.7041 and 0.1348, respectively. The ANN model employed the Levenberg-Marquardt algorithm with TrainLM and LearnGDM training functions, two hidden layers and one PEs for each hidden layer. Its performance regarding R^2 and RMSE values of 0.9299 and 0.0642 was highly acceptable. Apparently, with respect to the R^2 and RMSE, the ANN model was better than the ARIMA model.

Table 4 shows comparison between other works concerning different agricultural products. When comparing the R^2 and RMSE between ANN and ARIMA, the ANN showed a better R^2 than the ARIMA, and the RMSE of the ARIMA was higher than that for the ANN, like those in the forecasting of rainfall, predicting pod damage from pigeons [35,56–61]. While some of the agriculture predictions are favorable, the R^2 of ARIMA is better than the ANN model, such as predicting soil salt and water content in crop rootzones and prediction for sugarcane production in Bihar, etc. [62].

We aim to forecast the cherry coffee production of arabica coffee cultivated in northern Thailand. Two models in forecasting arabica coffee yields through ARIMA and ANN models were compared. The ARIMA model yielded a correlation coefficient (R^2) of 0.704 and an RMSE of 0.1348. The ANN model produced a higher R^2 of 0.9299 and a lower RMSE of 0.0642. In estimating yearly arabica coffee production, both models were determined to be adequate, but the ANN model appeared to perform better. However, when comparing the R^2 and RMSE with others in literature, shown in Table 4, it was found that the ANN and ARIMA models gave the reasonable R^2 and RMSE. They were suitable for coffee prediction.

With respect to the shortcomings of this work, they include missing data and the quality and quantity of data for coffee yield prediction. We considered merely six variable datasets; the area and productivity zone, rainfall, RH and temperature. The available amount of data remained low for these factors. Other factors that affect the coffee productivity, such as the amount of fertilizer, climate uncertainty each year, soil moisture, wind speed and amount of sunlight should also be considered, as they will help capture the full complexity of coffee yield. Moreover, flexible models that can capture the dynamic relationships between various factors affecting coffee yield may also be considered.

Table 4. Comparison of prediction performances with the literature.

Reference	Output	Period (yrs)	Predictors	R^2		RMSE	
				ANN	ARIMA	ANN	ARIMA
[56]	Chickpea production	5	rainfall, minimum and maximum temperatures	0.960	0.591	66.72	159.63
[59]	Wheat production	58	total annual precipitation, applied fertilizer, population and cultivated area	0.930	-	0.39	1.46
[62]	Soil salt and water content	5	crop rootzone	0.886	0.898	-	-
[57]	Behavioral pattern of rainfall	93	rainfall	0.984	0.953	5.518	35.88
[61]	Sugarcane production	81	area, production, yield	-	-	12.99	13.82
[58]	Crop planning	32	rainfall	0.790	0.750	93.97	97.12
[35]	Pod damage of pigeon pea	27	relative humidity	0.770	0.650	1.97	2.16
[60]	Agricultural and water resources	100	rainfall, temperature	-	-	59.03	76.78

For future works, application of other ML algorithms such as decision tree, random forest, support vector machine, K-nearest neighbors, K-mean clustering, principal component analysis, naive Bayes etc. may be considered. Other techniques such as data augmentation from multiple sources, sensitivity analysis and sustainability analysis may be incorporated. Moreover, the coffee prediction model may be combined with assessing the feasibility of using remote sensing data, such as satellite imagery, to supplement the existing predictor variables and improve the forecasting models. Factors affected by climate change may also be considered.

5. Conclusions

The productivity of arabica coffee varies depending on the cultivated area, total rainfall, ambient temperature and RH, among other factors. They affect the yield of cherry coffee in each month. Accurate forecast oof the crop yield is crucial in response to customer needs. We used ANN and ARIMA models to predict the yield of arabica coffee using time-series data from 2004 to 2018. It was shown that both models could forecast coffee production satisfactorily. Within the dataset considered, the ANN (R^2 and RMSE of 0.9299 and 0.0642) appeared to perform better than the ARIMA (R^2 and RMSE of 0.7041 and 0.1348) model.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was partially supported by Chiang Mai University. One of the authors (Y.K.) wishes to acknowledge the CMU Graduate School for Research Assistant grant. We also wish to thank the Supply Chain and Engineering Management Research Unit (SCEM), Chiang Mai University for providing research facilities. This research is part of the project "A Strategic Roadmap Toward the Next Level of Intelligent, Sustainable and Human-Centered SME: SME 5.0" from the European Union's Horizon 2021 research and innovation program under the Marie Skłodowska-Curie Grant agreement No. 101086487.

Author contributions

Conceptualization, K.Y.T. and N.T.; Methodology, Y.K. and N.T.; Data curation, Y.K.; Formal analysis, Y.K., A.B. and E.R.; Investigation, Y.K. and A.B.; Writing—original draft preparation, Y.K.; Writing—review and editing, N.T. and E. R.; Supervision, K.Y.T.; Funding acquisition, K.Y.T.

Data availability statement

The data from the Climate Department included rainfall, RH and minimum and maximum temperature. The Agricultural Economics Office and Meteorological Department provide the area and productivity zone.

Conflict of interest

All authors declare no conflicts of interest.

References

1. Food, Nations AOotU (2017) The future of food and agriculture: Trends and challenges: FAO.
2. Giovannucci D, Purcell T (2008) Standards and agricultural trade in Asia. *Soc Sci Res Netw Electron* J 34: 789–797. https://doi.org/10.2139/ssrn.1330266
3. Chittithaworn C, Islam MA, Keawchana T, et al. (2011) Factors affecting business success of small & medium enterprises (SMEs) in Thailand. *Asian Soc Sci* 7: 180–190. https://doi.org/10.5539/ass.v7n5p180
4. Anderson K (2022) Agriculture in a more uncertain global trade environment. *Agric Econ* 53: 563–579. https://doi.org/10.1111/agec.12726
5. Gu YH, Jin D, Yin H, et al. (2022) Forecasting agricultural commodity prices using dual input attention LSTM. *Agriculture* 12: 256. https://doi.org/10.3390/agriculture12020256

6. Sharafati A, Moradi Tayyebi M, Pezeshki E, et al. (2022) Uncertainty of climate change impact on crop characteristics: A case study of Moghan plain in Iran. *Theor Appl Climatol* 149: 603–620. https://doi.org/10.1007/s00704-022-04074-9
7. Somporn C, Kamtuo A, Theerakulpisut P, et al. (2011) Effects of roasting degree on radical scavenging activity, phenolics and volatile compounds of Arabica coffee beans (Coffea arabica L. cv. Catimor). *Int J Food Sci Technol* 46: 2287–2296. https://doi.org/10.1111/j.1365-2621.2011.02748.x
8. Haryono A, Maarif MS, Suroso A, et al. (2023) The design of a contract farming model for coffee tree replanting. *Economies* 11: 185. https://doi.org/10.3390/economies11070185
9. Azis AM, Irjayanti M, Rusyandi D (2022) Visibility and information accuracy of coffee supply chain in West Java Indonesia. In: Sergi BS, Sulistiawan D (Eds.), *Modeling Economic Growth in Contemporary Indonesia,* Emerald Publishing Limited, 225–236. https://doi.org/10.1108/978-1-80262-431-120221014
10. Katemauswa FA (2019) Factors influencing demand forecasting and demand planning: A case at an apparel retailer. MSc Dissertation, University of Kwazulu-Natal. https://researchspace.ukzn.ac.za/handle/10413/18966
11. Kilian B, Jones C, Pratt L, et al. (2006) Is sustainable agriculture a viable strategy to improve farm income in Central America? A case study on coffee. *J Bus Res* 59: 322–330. https://doi.org/10.1016/j.jbusres.2005.09.015
12. Kittichotsatsawat Y, Jangkrajarng V, Tippayawong KY (2021) Enhancing coffee supply chain towards sustainable growth with big data and modern agricultural technologies. *Sustainability* 13: 4593. https://doi.org/10.3390/su13084593
13. Kruse L, Wunderlich N, Beck R (2019) Artificial intelligence for the financial services industry: What challenges organizations to succeed. *Proceedings of the 52nd Hawaii International Conference on System Sciences,* 6408–6417. https://doi.org/10.24251/hicss.2019.770
14. Utku Al, Kaya SK (2022) Deep learning based a comprehensive analysis for waste prediction. *Oper Res Eng Sci: Theory Appl* 5: 176–189. https://doi.org/10.31181/oresta190822135u
15. Tanikić D, Manić M, Devedžić G, et al. (2010) Modelling metal cutting parameters using intelligent techniques. *J Mech Eng/Strojniški Vestnik*, 56: 52–62.
16. Agatonovic-Kustrin S, Beresford R (2000) Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J Pharm Biomed Anal* 22: 717–727. https://doi.org/10.1016/s0731-7085(99)00272-1
17. Liakos KG, Busato P, Moshou D, et al. (2018) Machine learning in agriculture: A review. *Sensors* 18: 2674. https://doi.org/10.3390/s18082674
18. Khairunniza-Bejo S, Mustaffha S, Ismail WIW (2014) Application of artificial neural network in predicting crop yield: A review. *J Food Sci Eng* 4: 1.
19. Kittichotsatsawat Y, Tippayawong N, Tippayawong KY (2022) Prediction of arabica coffee production using artificial neural network and multiple linear regression techniques. *Sci Rep* 12: 14488. https://doi.org/10.1038/s41598-022-18635-5
20. Bhojani SH, Bhatt N (2020) Wheat crop yield prediction using new activation functions in neural network. *Neural Comput Appl* 32: 13941–13951. https://doi.org/10.1007/s00521-020-04797-8
21. Palanivel K, Surianarayanan C (2019) An approach for prediction of crop yield using machine learning and big data techniques. *Int J Comput Eng Technol* 10: 110–118. https://doi.org/10.34218/ijcet.10.3.2019.013

22. Zhao Z, Chow TL, Rees HW, et al. (2009) Predict soil texture distributions using an artificial neural network model. *Comput Electron Agric* 65: 36–48. https://doi.org/10.1016/j.compag.2008.07.008
23. Kafy AA, Rahman AF, Al Rakib A, et al. (2021) Assessment and prediction of seasonal land surface temperature change using multi-temporal Landsat images and their impacts on agricultural yields in Rajshahi, Bangladesh. *Environ Challenges* 4: 100147. https://doi.org/10.1016/j.envc.2021.100147
24. Kaul M, Hill RL, Walthall C (2005) Artificial neural networks for corn and soybean yield prediction. *Agric Syst* 85: 1–18. https://doi.org/10.1016/j.agsy.2004.07.009
25. Abdollahpour S, Kosari-Moghaddam A, Bannayan M (2020) Prediction of wheat moisture content at harvest time through ANN and SVR modeling techniques. *Inf Proc Agric* 7: 500–510. https://doi.org/10.1016/j.inpa.2020.01.003
26. Ustaoglu B, Cigizoglu H, Karaca M (2008) Forecast of daily mean, maximum and minimum temperature time series by three artificial neural network methods. *Meteorol Appl* 15: 431–445. https://doi.org/10.1002/met.83
27. Tariq A, Yan J, Ghaffar B, et al. (2022) Flash flood susceptibility assessment and zonation by integrating analytic hierarchy process and frequency ratio model with diverse spatial data. *Water* 14: 3069. https://doi.org/10.3390/w14193069
28. Ghaderizadeh S, Abbasi-Moghadam D, Sharifi A, et al. (2022) Multiscale dual-branch residual spectral–spatial network with attention for hyperspectral image classification. *IEEE J Sel Topics Appl Earth Observ Remote Sens* 15: 5455–5467. https://doi.org/10.1109/jstars.2022.3188732
29. Zamani A, Sharifi A, Felegari S, et al. (2022) Agro climatic zoning of saffron culture in miyaneh city by using WLC method and remote sensing data. *Agriculture* 12: 118. https://doi.org/10.3390/agriculture12010118
30. Kosari A, Sharifi A, Ahmadi A, et al. (2020) Remote sensing satellite's attitude control system: Rapid performance sizing for passive scan imaging mode. *Aircr Eng Aerosp Technol* 92: 1073–1083. https://doi.org/10.1108/aeat-02-2020-0030
31. Pfaff B (2008) Analysis of integrated and cointegrated time series with R. Springer Science & Business Media. https://doi.org/10.1007/978-0-387-75967-8
32. Padhan PC (2012) Application of ARIMA model for forecasting agricultural productivity in India. *J Agric Soc Sci* 8: 50–56.
33. Iqbal N, Bakhsh K, Maqbool A, et al. (2005) Use of the ARIMA model for forecasting wheat area and production in Pakistan. *J Agric Soc Sci* 1: 120–122.
34. Osman T, Divigalpitiya P, Arima T (2016) Using the SLEUTH urban growth model to simulate the impacts of future policy scenarios on land use in the Giza Governorate, Greater Cairo Metropolitan region. *Int J Urban Sci* 20: 407–426. https://doi.org/10.1080/12265934.2016.1216327
35. Kumari P, Mishra G, Srivastava C (2017) Forecasting models for predicting pod damage of pigeonpea in Varanasi region. *J Agrometeorol* 19: 265–269. https://doi.org/10.54386/jam.v19i3.669
36. Bekuma T, Mamo G, Regassa A (2022) Modeling and forecasting of rainfall and temperature time series in East Wollega Zone, Western Ethiopia. *Arabian J Geosci* 15: 1377. https://doi.org/10.1007/s12517-022-10638-w

37. Mahto AK, Alam MA, Biswas R, et al. (2021) Short-term forecasting of agriculture commodities in context of indian market for sustainable agriculture by using the artificial neural network. *J Food Qual* 2021: 9939906. https://doi.org/10.1155/2021/9939906
38. Purohit SK, Panigrahi S, Sethy PK, et al. (2021) Time series forecasting of price of agricultural products using hybrid methods. *Appl Artif Intell* 35: 1388–1406.. https://doi.org/10.1080/08839514.2021.1981659
39. Cenas PV (2017) Forecast of agricultural crop price using time series and Kalman filter method. *Asia Pac J Multidiscip Res* 5: 15–21.
40. Onsree T, Tippayawong N (2021) Machine learning application to predict yields of solid products from biomass torrefaction. *Renewable Energy* 167: 425–432. https://doi.org/10.1016/j.renene.2020.11.099
41. Katongtung T, Onsree T, Tippayawong KY, et al. (2023) Prediction of biocrude oil yields from hydrothermal liquefaction using a gradient tree boosting machine approach with principal component analysis. *Energy Rep* 9: 215–222. https://doi.org/10.1016/j.egyr.2023.08.079
42. Prasertpong P, Onsree T, Khuenkaeo N, et al. (2023) Exposing and understanding synergistic effects in co-pyrolysis of biomass and plastic waste via machine learning. *Bioresour Technol* 369: 128419. https://doi.org/10.1016/j.biortech.2022.128419
43. Onsree T, Tippayawong N, Phithakkitnukoon S, et al. (2022) Interpretable machine-learning model with a collaborative game approach to predict yields and higher heating value of torrefied biomass. *Energy* 249: 123676. https://doi.org/10.1016/j.energy.2022.123676
44. Rahman MM, Islam MA, Mahboob MG, et al. (2022) Forecasting of potato production in Bangladesh using ARIMA and mixed model approach. *Sch J Agric Vet Sci* 10: 136–145. https://doi.org/10.36347/sjavs.2022.v09i10.001
45. Sankar TJ, Pushpa P (2022) Implementation of time series stochastic modelling for zea mays production in India. *Math Stat Eng Appl* 71: 611–621.
46. Nassiri H, Mohammadpour SI, Dahaghin M (2022) Forecasting time trends of fatal motor vehicle crashes in Iran using an ensemble learning algorithm. *Traffic Inj Prev* 24: 44–49. https://doi.org/10.1080/15389588.2022.2130279
47. Gorzelany J, Belcar J, Kuźniar P, et al. (2022) Modelling of mechanical properties of fresh and stored fruit of large cranberry using multiple linear regression and machine learning. *Agriculture* 12: 200. https://doi.org/10.3390/agriculture12020200
48. Salari K, Zarafshan P, Khashehchi M, et al. (2022) Modeling and predicting of water production by capacitive deionization method using artificial neural networks. *Desalination* 540: 115992. https://doi.org/10.1016/j.desal.2022.115992
49. Zhu X, Xiao G, Wang S (2022) Suitability evaluation of potential arable land in the Mediterranean region. *J Environ Manag* 313: 115011. https://doi.org/10.1016/j.jenvman.2022.115011
50. Wongchai W, Onsree T, Sukkam N, et al. (2022) Machine learning models for estimating above ground biomass of fast growing trees. *Expert Syst Appl* 199: 117186. https://doi.org/10.1016/j.eswa.2022.117186
51. Katongtung T, Onsree T, Tippayawong N (2022) Machine learning prediction of biocrude yields and higher heating values from hydrothermal liquefaction of wet biomass and wastes. *Bioresour Technol* 344: 126278. https://doi.org/10.1016/j.biortech.2021.126278
52. Pesaran MH (2007) A simple panel unit root test in the presence of cross-section dependence. *J Appl Econometrics* 22: 265–312. https://doi.org/10.2139/ssrn.457280

53. Suresh K, Krishna Priya S (2011) Forecasting sugarcane yield of Tamilnadu using ARIMA models. *Sugar Tech* 13: 23–26. https://doi.org/10.1007/s12355-011-0071-7
54. Eni D (2015) Seasonal ARIMA modeling and forecasting of rainfall in Warri Town, Nigeria. *J Geosci Environ Prot* 3: 91. https://doi.org/10.4236/gep.2015.36015
55. Sapna S, Tamilarasi A, Kumar MP (2012) Backpropagation learning algorithm based on Levenberg Marquardt Algorithm. *Comp Sci Inform Technol (CS and IT)* 2: 393–398. https://doi.org/10.5121/csit.2012.2438
56. Rawat S, Mishra AR, Gautam S, et al. (2022) Regional time series forecasting of chickpea using ARIMA and neural network models in central plains of Uttar Pradesh (India). *Int J Environ Clim Change* 2022: 2879–2889. https://doi.org/10.9734/ijecc/2022/v12i1131280
57. Somvanshi V, Pandey O, Agrawal P, et al. (2006) Modeling and prediction of rainfall using artificial neural network and ARIMA techniques. *J Ind Geophys Union* 10: 141–151.
58. Dwivedi D, Kelaiya J, Sharma G (2019) Forecasting monthly rainfall using autoregressive integrated moving average model (ARIMA) and artificial neural network (ANN) model: A case study of Junagadh, Gujarat, India. *J Appl Nat Sci* 11: 35–41. https://doi.org/10.31018/jans.v11i1.1951
59. Latifi Z, Shabanali Fami H (2022) Forecasting wheat production in Iran using time series technique and artificial neural network. *J Agric Sci Technol* 24: 261–273.
60. Sekhar PH, Kesavulu Poola K, Bhupathi M (2020) Modelling and prediction of coastal Andhra rainfall using ARIMA and ANN models. *Int J Stat Appl Math* 5: 104–110.
61. Paswan S, Paul A, Paul A, et al. (2022) Time series prediction for sugarcane production in Bihar using ARIMA & ANN model. *The Pharma Innovation J* 11: 1947–1956.
62. Zou P, Yang J, Fu J, et al. (2010) Artificial neural network and time seriesmodels for predicting soil salt and water content. *Agric Water Manag* 97: 2009–2019. https://doi.org/10.1016/j.agwat.2010.02.011

AIMS Press