
















Shallot Crop Harvest Time and Yield Prediction Using Machine Learning Based on Farmers' Tacit Knowledge in Brebes Regency, Indonesia

Arief Arianto¹, Gadang Ramantoko², Agung Hendriadi^{1*}, Maya Ariyanti², Noveria Sjafrina¹,
Boni Benyamin¹, Helni M. Jumhur², Yogi Purna Rahardjo¹, Huda M. Elmatrani¹, Puji Astuti¹,
Ratri Wahyuningtyas², Mulyanto Mulyanto¹, Mulyana Hadipernata¹

¹ Research Center for Agroindustry, National Research and Innovation Agency, Tangerang Selatan 15314, Indonesia

² School of Economics and Business, Telkom University, Bandung 40257, Indonesia

Corresponding Author Email: agun044@brin.go.id

Copyright: ©2025 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ij dne.200321>

ABSTRACT

Received: 4 September 2024

Revised: 25 November 2024

Accepted: 6 January 2025

Available online: 31 March 2025

Keywords:

farmer experiences, harvest time, machine learning, productivity, tacit knowledge, yield prediction

Shallot price fluctuations in Indonesia are caused by a lengthy supply chain and limited production, which makes supply control challenging. This study employed machine learning to forecast shallot yields and harvest times in Brebes Regency, Central Java, one of the major production areas. Data were collected through farmer interviews, which encompassed productivity and farming practices, and analyzed using twelve machine learning algorithms, including Gradient Boosting, AdaBoost, XGB, ElasticNet, and Decision Trees. Model performance was evaluated using MSE, MAE, and R-squared values, with ElasticNet being identified as the most accurate. Harvest time predictions were influenced by plant age and morning temperature, while yield depended on factors such as planted area, bed dimensions, daily temperature range, bulb weight, and phosphorus levels. Farmers' tacit knowledge was also incorporated, improving the model's reliability. The deployment results revealed a 13% deviation between predicted and actual yields, demonstrating reasonable accuracy. However, the error margin for harvest time predictions was 23.5%, reflecting the complexity of environmental and operational factors. The study provides a data-driven framework for understanding shallot productivity and the variables influencing it, and offers insights into improving forecasting models for more effective agricultural planning.

1. INTRODUCTION

Indonesia has a high demand for shallots, but seasonal production variations cause price fluctuations. The market is asymmetrical, as a significant increase in consumer-level demand directly impacts production centers. Conversely, supply shortages disproportionately impact farmers' prices, driven by high trade and transportation margins resulting from an extended supply chain [1]. Brebes Regency, in Central Java, is a shallot production center that supplies 60% of shallots to the Jakarta market. In Brebes, substantial changes in harvest area and production directly affect the sustainability of shallot farming. Market price information is easily accessible from Jakarta's market.

Farmers often rely on the Ijon system, a pre-harvest selling arrangement that provides immediate financial relief but can cause them to lose control over their harvest and miss potential price increases. This issue highlights the need for precise yield forecasting to empower farmers in negotiating equitable repayment terms and achieving better financial outcomes. Data analysis and machine learning can help estimate crop yields accurately by integrating farmers' experience with environmental and production variables [2].

Machine learning-based prediction models play a critical role in effective crop farming, assisting in decisions related to

planting, irrigation, fertilization, harvesting, and trading [3]. Among the machine learning approaches, the MARS-ANN hybrid model demonstrates high prediction accuracy by combining ANN's predictive power with MARS's feature selection capabilities. For example, the model has been used effectively to predict wheat, rice, and maize yields based on meteorological and soil data [4]. Similarly, machine vision-based yield monitoring has been employed to create geotagged yield maps for shallot fields, achieving a 76% detection accuracy [5]. In Turkey, onion yield predictions utilized support vector regression and polynomial regression, while in Bangladesh, climatic data was combined with linear regression for shallot yield estimations [6, 7]. Furthermore, the SVM classifier has been used to assess shallot quality with 60% accuracy, highlighting its potential for integration into web or mobile tools [8].

The effectiveness of machine learning-based yield prediction can be evaluated across three dimensions: prediction horizon, scale, and crop type [9]. Predictions are feasible at all vegetation stages, but many studies focus on predictions right before harvest. The best grain forecasts for each model were made before harvest at the start, middle, and end of the growing season [10]. Scale is important because models at each scale serve different purposes, such as plant-scale models aiming to understand factors affecting crop

growth, field-scale models assisting in crop management [11], and models at larger scales primarily informing policy-making in agriculture. Klompenburg's "Crop Prediction" model is a comprehensive framework that integrates 21 factors to enhance yield predictions on various scales [12].

Although machine learning has been extensively studied for crop yield prediction, its ability to integrate tacit knowledge, such as farmers' intuition, experience, and observational skills, has not been thoroughly explored [13]. Tacit knowledge is qualitative and context-specific, often communicated through actions rather than words. Designing systematic yet simple questionnaires that elicit detailed observations from farmers can help encode this knowledge into machine learning processes, enhancing model accuracy and relevance.

This study directly addresses the challenges Indonesian shallot farmers face, such as market dynamics and the Ijon system, by leveraging machine learning for yield and harvest time predictions. Based on the discussion above, this paper aims to: 1) identify the most effective machine learning model for predicting shallot crop yield and harvest time, and 2) determine the critical variables, including tacit knowledge, that influence these predictions.

2. THE COMPREHENSIVE THEORETICAL BASIS

2.1 Soil conditions and fertilizer

Klompenburg's 'Crop Prediction' model used 21 factors to enhance yield predictions, including soil properties such as pH, moisture, and texture, as well as weather conditions like temperature, rainfall, and humidity [12]. Shallots prefer slightly acidic soil with higher CEC, as it retains nutrients for roots and increases the adsorptive capacity for cations in soil [14]. Crop yields were often predicted using nutrients in the soil, NDVI, and meteorological components [15]. The physical size of the tubers has a strong correlation with the leaf area index, which is an important element in growth and nutrient uptake [16].

2.2 Lighting conditions

Crop productivity can also be impacted by environmental variables like wind speed and sunshine exposure [17]. The development of shallots is also influenced by lighting other than sunlight. Their growth is influenced by two light mechanisms: photoperiod and gamma radiation. It has been demonstrated that controlled exposure to low levels of gamma radiation stimulates bulb development in shallots by inducing hormonal changes that expand bulb size and potentially have an impact on yield [18]. On the other hand, the amount of light exposure each day that impacts the shallot growth cycle is known as photoperiod. The best period of daily light exposure is when around 70% of the sunshine exposure occurs during the day [18].

2.3 Watering needs

Shallot plants require frequent watering in the beginning, followed by daily irrigation during growth, and then less watering for bulb formation [19, 20]. Shallots can use up to 81.17% of the available water for evapotranspiration [21]. Effective water management also depends on ridge size, width, and row layout to avoid waterlogging, which can lead to bulb

rot and other challenges [22].

2.4 Cultivation strategies

The Klompenburg model also considers agronomic practices like fertilizer application, irrigation methods, planting density, crop characteristics like variety/genotype, growth stage timing, pest and disease incidence, and disease outbreaks [12]. Planting strategies such as increasing planting density, variety selection, planting distance, and bed size may contribute to higher shallot production. Increasing planting density must be carefully managed with a suitable planting distance and supported by other agronomic methods so that growth does not become restricted and harvest outcomes are maximized [23]. Excessive density will cause shallot plants to compete with one another for nutrients, water, and sunshine, which will stunt their growth and cause their bulbs to shrink. Insufficient air circulation will also increase the plants' vulnerability to disease attacks [24].

2.5 Farmer knowledge

Farmers' tacit knowledge, derived from years of experience and local adaptations, is invaluable for tailoring practices to specific conditions. Farmers select shallot varieties depending on market demand, pest and disease resistance, and production levels. The Bima Brebes shallot variety has huge bulbs, disease resistance, and a strong flavor [25]. However, in Tapin Regency, many farmers have ceased planting shallots due to the adverse effects of climate change, including stagnant water, fruit, and root rot, and decreased harvest and sale prices [26]. Farmers' understanding of optimal harvest timing, typically at 55-60 days post-flowering, and their ability to adapt practices contribute significantly to crop success [27]. Moreover, their knowledge helps maintain family food security and achieve high farming success rates, as seen in Malumbi Village, where motivation and competence correlate strongly with outcomes [28]. This tacit knowledge becomes especially critical when farmers face challenges such as low market prices. For instance, in Brebes, the lower price limit for shallots is IDR 13,730.49, with a fluctuation coefficient of 0.20, indicating a low-risk scenario [29].

3. METHOD

3.1 Experimental site

The study was conducted in Brebes, in the Central Java province of Indonesia (7°3'0" S, 108°54'0" E). This area has a tropical climate with temperatures ranging from 24°C to 32°C within the two months, with the lowest temperatures from June to August. The average annual rainfall is around 1200-2000 mm/year. In Brebes Regency, rain can occur in any month, although seasonal rainfall typically begins at the end of October and lasts until May. The probability of rain in a month is more than 80% in January and February.

3.2 Data collection

Data were collected via farmer interviews using questionnaires with the assistance of Brebes district extension personnel in nine sub-districts within the Brebes district, specifically Brebes, Wanasari, Ketanggungan, Larangan,

Bulakamba, Kersana, Jatibarang, Bantarkawung, and Bantarharjo. These sub-districts were selected as the main shallot production areas. The diversity in agroclimatic conditions, soil properties, and farming practices across these areas ensures a comprehensive dataset for modeling shallot yield and harvest time.

The dataset gathered from Brebes' extension staff totaled 368, with the condition that the shallot plants were 45 days old when the samples were collected. This age was chosen as it is a critical growth phase where vegetative development and early bulb formation occur, allowing for accurate observation of key yield predictors.

Three samples were taken from each farmer's field to ensure representative and reliable data. This approach minimizes the impact of within-field variability caused by uneven soil properties, microclimatic differences, or irrigation patterns, thereby improving the robustness of the data. The sample size

and distribution were designed to capture sufficient variability while maintaining feasibility for field data collection.

Tools such as rulers, scales, and vernier calipers are used to measure growth parameters. Information on the use of fertilizer in shallot plants was recorded. We also used smartphone weather applications to obtain climate data, including temperature, humidity, and precipitation. Agricultural extension conducts data gathering to minimize measurement mistakes. Farmers were then interviewed again after thirty days following the first interview to collect data on the harvest day and yields. All data was collected and stored in Microsoft Excel. They also take note of each field's soil type, texture, and pH. Farmers were also asked about how much water is consumed and how to water it. Several questions related to the condition of plant area, bed size, cultivation methods and farmers' tacit knowledge are detailed in the questionnaire in Table 1.

Table 1. The questions for farmer respondents

Klompenburg Data Structure	Questions	Data Type	Unit
Crop Information	How old will your shallot plants be when harvested?		days
	What is the height of a shallot plant at 45 days old?	Main; weighed;	cm
	How many leaves will there be on a 45-day-old shallot plant?	measured;	leaves
	How many shallot bulbs will there be at 45 days old?	calculated	bulb
	What was the physical size of the tubers and the number of tubers per clump after 45 days?		cm
Leaf Area Index	What size is the shallot plant leaf area at 45 days old?	Primary; measured	cm ²
Soil Type	What is the soil type?	Primary	cat.
pH Soil	What is the pH of the growing medium?	Primary; measured	value
Rainfall	Does the amount of rainfall affect the productivity and yield of shallots, and is the availability of water in the fields sufficient?	Secondary	cat.
Wind Speed	What is the wind speed in a day?	Primary	m/s
Humidity	What is the average humidity in a day?	Primary; measured	%
Nutrients in Soil	What types of nutrients in the soil are needed for plant shallot growth?		
	How much is the percentage of N, P, and K?	Secondary	%
Irrigation	How much watering is needed per day at each stage of the shallot plant?	Primary	cat.
Fertilization	What is the dose of fertilizer for plant development and shallot bulb yield?		
	What is the fertilizer requirement in percentages of N, P, and K?	Primary	kg
Temperature	What is the morning temperature in a day?	Primary	°C
Variety	What variety of shallots are currently being planted on the land?	Primary	cat.
Land Size	What is the current area of shallot land planted?	Primary	m ²
Planting Space	What is the planting distance?	Primary	m ²
Width and Number of Beds -Planting Area	What is the current width and length of the shallot bed?		
	How many beds are there in the area you are currently planting?	Primary	beds
Elevation	How high is your shallot field? Calculating the height can be helped by using the extension officer's smartphone.	Primary	m
Weed, Pest, and Diseases	What is the weed infestation rate?		
	Do you carry out pest and disease control regularly?	Primary	%
Farmer's Tacit Knowledge			
Rainfall	Is the quantity of rain sufficient?	Primary	cat.
Reason for Planting	What are the considerations for planting?	Primary	cat.
	If shallot prices are low, is the farmer still considering planting?	Primary	cat.
	Estimated price at harvest time: is it profitable?		
Price	What is the price of the product when planting?	Primary	cat.
Harvest Time	Do you hurry up the harvest, and how long does it take to speed up if prices are high as the harvest gets close?	Primary	cat.
Wind	Is there any benefit to the wind blowing faster this month?	Primary	cat.
Motivation	Are you confident that you will get good results this planting season?	Primary	cat.

3.3 Dataset enhancement

The acquired data is then prepared and segregated into numerical and categorical data. The data underwent cleaning and pre-processing, including tasks such as encoding category data, normalizing numerical data, and scaling the data. Subsequently, both quantitative and qualitative data are

merged. The dataset of 368 obtained when used directly in the machine learning process is still relatively small, so additional data is needed through data synthesis activities. The data synthesis process was conducted to replicate real-world conditions of shallot farming, ensuring the synthetic data retained the statistical and contextual integrity of the original dataset. Machine learning models require large volumes of

data to be trained and validated. By using large data sizes, they can find meaningful patterns in real-life data [30]. It is known that most conventional machine learning methods produce good accuracy results when the dataset has high dimensionality [31].

The synthesis process involved the following steps:

1. Data preprocessing: The real data was first cleaned and analyzed for patterns, distributions, and correlations to understand the underlying statistical properties.

2. Synthetic data generation using SDV-CTGAN. The Synthetic Data Vault (SDV) library, in combination with the Conditional Tabular Generative Adversarial Network (CTGAN), was utilized. CTGAN is particularly effective in generating realistic tabular data with a small dataset, as it captures the distribution and dependencies between features. The model was trained iteratively, adjusting hyperparameters to achieve replication values above 75%.

3. Validation of synthetic data. The synthetic dataset was evaluated by comparing statistical properties such as mean, variance, and feature correlations with the original data. This step confirmed that the synthetic data retained the characteristics of shallot farming conditions.

4. Augmentation using Gretel AI. A cloud-based synthetic data platform was used to complement SDV-CTGAN. This platform employed advanced algorithms to generate additional synthetic data, offering a user-friendly interface and API for seamless integration. It also provided tools to compare synthetic data against real data for quality assurance.

5. Final Merging. The validated synthetic data was merged with the original dataset, creating a more comprehensive dataset for machine learning. This combined dataset was then split into training, validation, and test sets to ensure robust model evaluation.

3.4 Predictive modeling

A crucial step in machine learning is analyzing algorithm models using the train-test procedure. The machine learning process for estimating harvest time and yield of shallot crops consists of three phases: optimizing the hyperparameters of the base estimator model, optimizing the hyperparameters of the competitor model, and selecting the superior estimator between the base model and the competitor model. The training process uses synthetic data with 12 model algorithms, according to Figure 1.

The next step involves applying machine learning algorithms to the modeling process. We experimented with several methods to achieve improved performance. The algorithms used in this study included FNN, AdaBoost, XGB, SVR, Lasso, Ridge CV, decision tree, gradient boosting, ElasticNet, extra trees, and linear regression. Both hyperparameter tuning and optimum model selection utilize evaluation metrics given by the Sklearn library. Various metrics are utilized to assess the reliability of the optimal model across multiple tests. Based on 70% of the synthesis dataset, we trained the models to produce several alternative models, and the remaining 30% of the data was used in the validation phases. The models are compared using a mean square error (MSE) and a mean absolute error (MAE) to see which performs the best and has the lowest MAE and MSE values. Furthermore, the biggest R-squared value is also evaluated. All data sets without synthesis are reused in the testing and validation process.

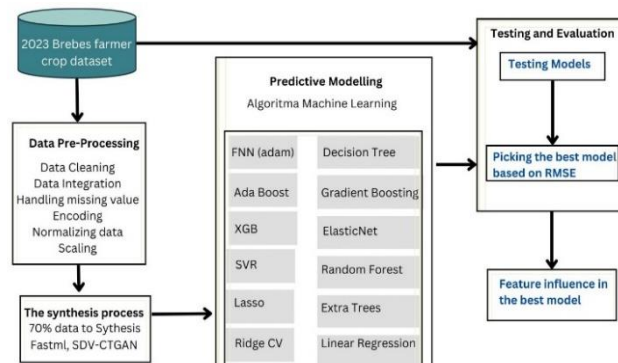


Figure 1. Machine learning workflow was used in this study

3.5 Feature selection to coefficient model

The best prediction model can perform feature selection during the model fitting process. This means it can identify and potentially remove irrelevant or redundant features that might not contribute significantly to predicting yield or harvest time. The best prediction model shrinks coefficients of less important features and eliminates some, resulting in a more interpretable model. This reduces model complexity, enhances prediction accuracy, and ensures efficient models. This tool is used in machine learning models to identify important characteristics that determine shallot yield and harvest time, as well as key tacit knowledge variables.

3.6 Feature selection to coefficient model deployment machine learning model into Google spreadsheet

The model acquired from the training and testing analysis in the machine learning process in the form of linear programming is then entered into a Google spreadsheet with the code "`= array formula (if (row (B:B) = 1)); formula calculation`". After the farmers' data is input, all variables are normalized. This normalization method involves scaling the input variables to a specific range, often 0 to 1, which helps alleviate difficulties associated with varied input feature scales and improves the model's numerical stability. Each input's value is calculated using the linear programming equation, which includes categorical data. The model calculation results are transformed with the resulting data's normalized range value to provide the desired prediction results.

The model that was deployed in the Google spreadsheet was then evaluated on 15 shallot farmers who were interviewed, and samples of shallot plants were collected at the age of 45 days as in method 3.2 and then interviewed again following the first interview to discuss the harvest day and results. The data supplied into Google spreadsheet was then calculated and reported via an email sent by the system. The computation results must be compared to the data from farmer interviews to determine the difference in error.

4. RESULTS AND DISCUSSION

4.1 Verification structure

The crop forecasting model proposed by Klompenburg divides several groups of 21 variables that are arranged like groups of latent variables. The grouped latent variables can be analysed using factor analysis techniques. The Kaiser-Meyer-

Olkin (KMO) and Bartlett tests are commonly used to assess whether factor analysis requirements can be performed on a set of numerical data. The KMO test, which ranges from 0 to 1, tests the fit of each observed variable as well as the overall model. Bartlett's sphericity test tests whether the observed variables are correlated.

The results of the KMO test are 0.66 and the Bartlett test values are 245784.0 and 0.0. In the Bartlett test with a highly significant p-value (0.0), we can suggest the null hypothesis

(all variances are identical) can be rejected. Both the KMO (borderline) and Bartlett's tests show that the dataset is suitable for factor analysis. Figure 2 shows the results of the factor analysis of the data set, which shows the existence of factor grouping with green to yellow cells. However, the number of latent variables formed is not as many as those grouped by Klompenburg. Some of the remaining cells do not form groups and remain unique.

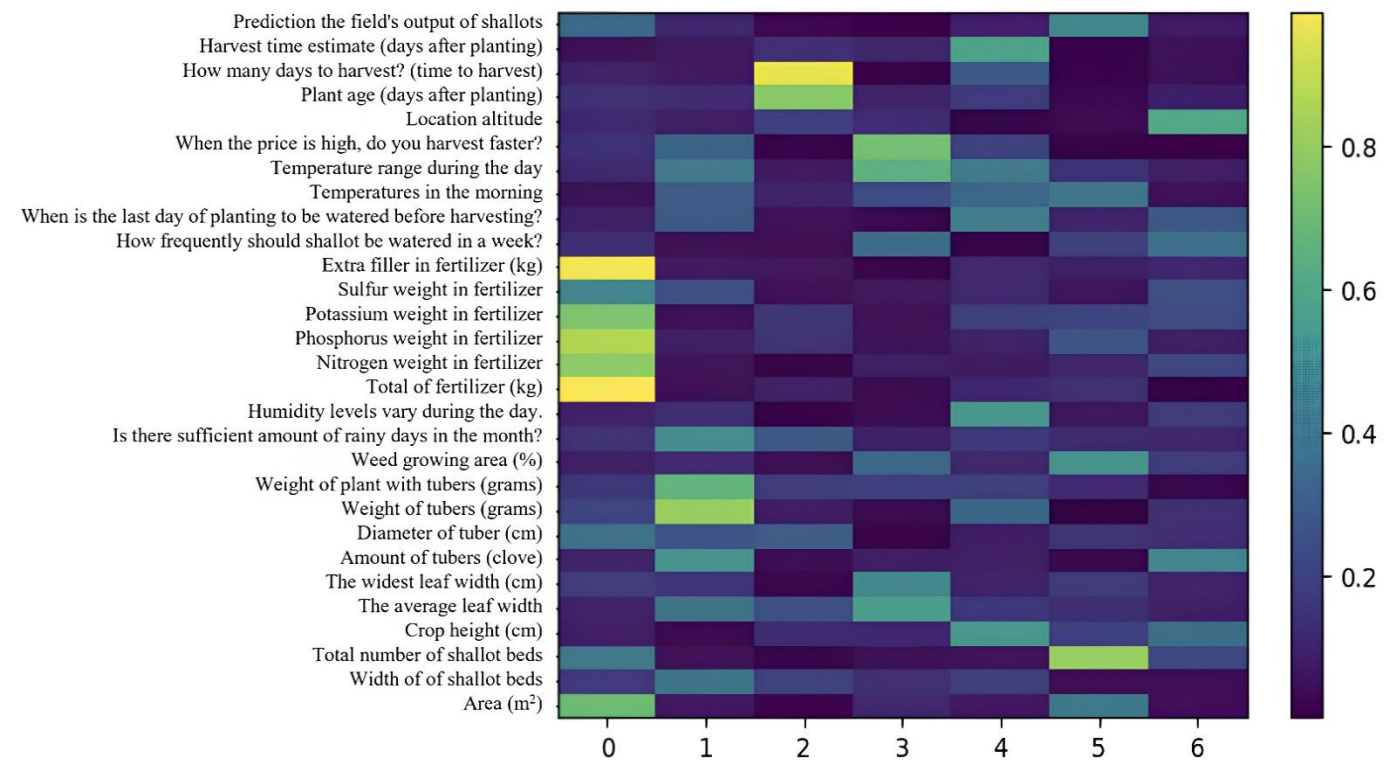


Figure 2. Factor loading of the dataset using varimax rotation having numerical part of dataset

4.2 Best model selection

The score values from MAE, MAPE, MSE, and R-squared were taken into consideration during a prediction model analysis to determine the best prediction value, as shown in Table 2. RidgeCV is the best model when considering the mean absolute error, mean squared error, and R-squared scores combined. However, several linear regression models, such as ElasticNet and Lasso, predict TTH and yield with negligible variations in the outcomes. ElasticNet was chosen as the best prediction model by a different method that employed the best estimator application. Combining L1 and L2 penalties from the Lasso and Ridge models, ElasticNet is a linear regression model. This model can successfully handle data with many associated characteristics, preventing overfitting and preserving model stability. When applied to various and complicated data sets, the Elasticnet model performs better than the Lasso model in the event of outlier data, resulting in more consistent and dependable findings. This makes ElasticNet more consistent and reliable in various data sets. In addition, Table 2 also demonstrates that the model predicts TTH better than yield.

4.3 Influence factor of yield and time to harvest

The best prediction model chosen is the Elasticnet model. ElasticNet may arrange the regression coefficient values,

which are shown in Table 3, from largest to smallest to perform feature selection during the model fitting process. On the left side of the table, the time to harvest (TTH) regression coefficients are presented in ascending order, and on the right side, the Yield is displayed in decreasing order.

Among the most significant predictors for TTH, plant age prediction (HST) and morning temperature range stand out. The negative coefficient for plant age prediction suggests that farmers' ability to predict optimal harvest times, based on experience and growth pace, is critical. A negative coefficient value indicates that the prediction of harvest time is getting closer to plant age (HST).

Morning temperature, a proxy for daily heat accumulation, correlates with physiological processes in plants, aligning with the heat unit method [32]. These two factors interact synergistically: consistent morning temperatures provide reliable data for growth rate estimation, enhancing the accuracy of farmers' harvest predictions. Additionally, the influence of rainfall, moisture levels, and soil type indicates the interplay between climatic and edaphic factors. While rainfall contributes to water availability, excessive amounts can delay harvest by prolonging vegetative growth. The morning temperature's moderation of daily heat accumulation further supports timely harvesting under variable rainfall conditions.

In Table 3, the yield predictors include planting area, bed structure (number and width), and day temperature range,

emphasizing the interplay between agronomic practices and environmental conditions. The planting area is also influenced by the number of beds and the width of the beds and affects the density of plants in one area [33].

Larger planting areas and optimized bed dimensions influence plant density, which can enhance sunlight exposure

and nutrient uptake. The day temperature range, with its positive coefficient, underscores how moderate fluctuations promote metabolic activities that boost bulb size and quality [34].

Table 2. Comparison of model's performance because of best model selection

Score		MAE	MSE	MAPE	R-squared
FNN (adam)	TTH	0.0773	0.0096	1.8502E-13	0.7225
	Yield	0.0994	0.0156	4.1366E-13	0.4845
	Overall	0.0884	0.0126		0.6035
Decision Tree Regressor	TTH	0.0985	0.0157	3.1768E-13	0.5475
	Yield	0.1275	0.0254	7.4422E-13	0.1574
	Overall	0.1130	0.0206		0.3525
Ada Boost Regressor	TTH	0.0892	0.0122	5.3668E-13	0.6477
	Yield	0.1138	0.0187	9.9146E-13	0.3814
	Overall	0.1015	0.0155		0.5146
Gradient Boosting Regressor	TTH	0.0791	0.0098	2.4870E-13	0.7177
	Yield	0.0995	0.0153	5.5496E-13	0.4938
	Overall	0.0893	0.0125		0.6136
XGB Regressor	TTH	0.0821	0.0106	2.8848E-13	0.6956
	Yield	0.1041	0.0164	6.4942E-13	0.4552
	Overall	0.0931	0.0135		0.5837
SVR Regressor	TTH	0.0774	0.0094	2.5630E-13	0.7293
	Yield	0.0981	0.0146	5.7628E-13	0.5162
	Overall	0.0877	0.0120		0.6301
ElasticNet	TTH	0.0773	0.0094	2.5419E-13	0.7304
	Yield	0.0979	0.0146	5.6932E-13	0.5164
	Overall	0.0875	0.0120		0.6309
Lasso	TTH	0.0770	0.0093	2.5313E-13	0.7310
	Yield	0.0979	0.0146	5.6525E-13	0.5164
	Overall	0.0875	0.0120		0.6312
Ridge CV	TTH	0.0770	0.0093	2.5361E-13	0.7312
	Yield	0.0979	0.0146	5.6645E-13	0.5164
	Overall	0.0874	0.0120		0.6313
Random Forest Regressor	TTH	0.0834	0.0109	3.3871E-13	0.6871
	Yield	0.1069	0.0172	7.1030E-13	0.4315
	Overall	0.0874	0.0140		0.5682
Extra Trees Regressor	TTH	0.0830	0.0106	3.2207E-13	0.6934
	Yield	0.1068	0.0171	7.2378E-13	0.4325
	Overall	0.0949	0.0139		0.5721
Linear Regression	TTH	0.0770	0.0093	2.5313E-13	0.7310
	Yield	0.0979	0.0146	5.6525E-13	0.5164
	Overall	0.0875	0.0120		0.6312

Table 3. Regression coefficient list produced by tuned ElasticNet regressor

Model Coefficient for Time to Harvest	Coefficient	Model Coefficient for Yield	Coefficient
<i>Plant Age Prediction (day after plant)</i>	-0.846	Area (m ²)	0.520
Morning Temperature Range	-0.139	Number of Beds	0.293
Rainfall amount in the last month	-0.082	Bed Width	0.167
Moisture Value (%)	-0.078	Day Temperature Range	0.146
Paint-Type of Soil (sand, colour, clay, crumb, hard)	-0.060	Bulb Weight (gram)	0.143
Bed Width	-0.056	Phosphor Weight (kg)	0.131
Tuber Diameter (cm)	-0.050	Plant Height (cm)	0.084
Weed Grown Area (%)	-0.050	Number of Tubers (Cloves)	0.055
Nitrogen Weight (kg)	-0.050	How many times a week watered?	0.044
Potassium Weight (kg)	-0.047	<i>Watering Methods</i>	0.038
Area (m ²)	-0.040	<i>Estimated harvest during high price period (How many days earlier?)</i>	0.036
Last Watering Close to Harvest (day after plant)	-0.039	Number of Leaves (Sheet)	0.033
Location Altitude	-0.037	Sulphur Weight (kg)	0.028
<i>Estimated Harvest During High Price Period (How many days earlier?)</i>	-0.037	Moisture Value (%)	0.028
Number of Tubers (Cloves)	-0.025	<i>Plant Age Prediction (day after plant)</i>	0.028
<i>What determines the start of the season (Price, weather, pest infestation, planting)?</i>	-0.021	<i>What determines the start of the season (price, weather, pest infestation, planting)?</i>	0.019

Model Coefficient for Time to Harvest	Coefficient	Model Coefficient for Yield	Coefficient
<i>Watering Methods</i>	-0.020	<i>Paint-Are you confident of good yields this planting season?</i>	0.016
<i>Using POC/Mo/Soil Improver</i>	-0.019	<i>Morning Temperature Range</i>	0.016
<i>Paint- Are you confidence of good yields this planting season?</i>	-0.011	<i>Paint-Plant Spacing</i>	0.015
<i>Weight of Plant with Tuber (gram)</i>	-0.003	<i>Paint-Type of Soil (Sand, colour, clay, crumb, hard)</i>	0.013
<i>If the price is low will you plant?</i>	0.000	<i>Weed grown area (%)</i>	0.000
<i>Is the weather favourable for the crop?</i>	0.000	<i>Total Fertilizer (kg)</i>	0.000
<i>Is the wind blowing faster this month useful?</i>	0.000	<i>Last watering close to harvest (day after plant)</i>	0.000
<i>Total Fertilizer (kg)</i>	0.000	<i>Is the wind blowing faster this month?</i>	0.000
<i>Widest Leaf Width (cm)</i>	0.000	<i>Is the wind blowing faster useful?</i>	0.000
<i>Whether there are many weeds this planting season?</i>	0.000	<i>Is the weather favourable for the crop</i>	0.000
<i>Paint-Plant Spacing</i>	0.002	<i>Is the amount of rain sufficient in the past two months?</i>	-0.002
<i>Is the wind blowing faster this month?</i>	0.002	<i>Using POC/Mo/soil improver</i>	-0.004
<i>Number of Beds</i>	0.002	<i>Widest Leaf Width (cm)</i>	-0.005
<i>Bulb Weight (gram)</i>	0.014	<i>If the price is low will you plant?</i>	-0.006
<i>The expected soil condition after watering</i>	0.025	<i>The expected soil condition after watering</i>	-0.006
<i>How many times a week watered?</i>	0.026	<i>Whether there are many weeds this planting season?</i>	-0.032
<i>Number of leaves (Sheet)</i>	0.029	<i>Location Altitude (masl)</i>	-0.039
<i>Phosphor Weight (kg)</i>	0.040	<i>Weight of plant with tuber (gram)</i>	-0.074
<i>Is the amount of rain sufficient in the past two months?</i>	0.040	<i>Potassium Weight (kg)</i>	-0.082
<i>Plant Height (cm)</i>	0.052	<i>Rainfall amount in the last month</i>	-0.085
<i>Other Weight (kg)</i>	0.065	<i>Other Weight (kg)</i>	-0.090
<i>Sulphur Weight (kg)</i>	0.077	<i>Tuber Diameter (cm)</i>	-0.118
<i>Day Temperature Range</i>	0.086	<i>Nitrogen Weight (kg)</i>	-0.146

Balanced fertilization plays a critical role in maximizing shallot yields, with phosphorus standing out as a key nutrient. Phosphorus significantly impacts yield by fostering healthy root systems, which enhance the plant's ability to efficiently absorb water and nutrients from the soil. This nutrient is also integral to various metabolic processes that convert energy and support plant growth, ultimately leading to improved yields. While nitrogen and potassium are essential for overall growth, their lower regression coefficients suggest that balanced fertilization holds greater importance than the dominance of individual nutrients.

Interestingly, the absence of potassium (K) does not substantially reduce shallot yields, whereas deficiencies in nitrogen, phosphorus, magnesium, or sulfur can lead to a noticeable decrease in bulb dry weight [35]. Adequate phosphate levels, therefore, not only promote robust root development but also ensure the efficient execution of metabolic functions critical for plant health and productivity. This reinforces the need for a balanced approach to fertilization, where each nutrient plays a synergistic role in achieving optimal yields.

4.4 Statistical evidence supporting factor importance

The ElasticNet model assigns coefficients based on their predictive power, removing redundant or negligible variables (coefficients = 0). For example, total fertilizer and wind conditions have zero coefficients in both TTH and Yield models, suggesting limited direct impact. Conversely, bulb weight and bed width consistently exhibit positive coefficients, underscoring their critical roles in Yield predictions. These statistical insights reinforce the importance of prioritizing key agronomic and environmental factors.

Table 3 outlines tacit knowledge components in the machine learning model related to harvest time and yield prediction. For harvest time, key factors include 1) prediction of harvest time, 2) day length (which increases when product

prices are high), 3) reasons for planting season timing, and 4) confidence in crop yields. Experienced farmers can predict harvest timing using plant indicators, weather data, and historical trends. In the yield section, factors such as day length, harvest time prediction, planting season reasons, confidence in harvest results, and weed attacks influence the model. Watering during planting has a greater influence on yield than planting distance, with coefficients of 0.038 and 0.015, respectively.

The comparison of regression models (linear regression and random forest) in Table 4 shows that while average predictions are similar, standard deviations vary. No model matches the standard deviation of the given test, and the random forest regressor weakly maintains minimum and maximum values. This comparative analysis highlights the strengths and weaknesses of each method, guiding the selection of the most suitable model.

4.5 Model performance and qualitative discussion

The model calculations that have been implemented on Google spreadsheet are then compared with the actual results at harvest age and production results shown in Table 5. The ElasticNet model achieves a TTH error deviation of approximately 1.9 days (23.5%) and a yield error of 166.9 kg/m² (13%), with prediction accuracy exceeding 80% as demonstrated in the study of Apriyanti et al. [36], which used feature extraction for orchid identification. However, these numerical results should be contextualized with on-ground realities. Farmers value timely and accurate predictions, especially when aligning harvests with market demands. For instance, predicting harvest timing during high-price periods can significantly enhance economic returns despite a minor trade-off in prediction accuracy.

Qualitatively, farmers express satisfaction when models reduce uncertainty, particularly in dynamic climates. However, challenges like yield variability due to unforeseen weather

changes or suboptimal input availability highlight areas for improvement. These insights suggest that while the model is

robust, integrating real-time environmental monitoring and farmer feedback loops could further enhance its utility.

Table 4. Comparison of statistical description prediction model produced by linear regression and random forest regressor

Stats.	Data for Testing		Linear Regression		Random Forest Regressor	
	TTH (days)	Yield (kg)	TTH (days)	Yield (kg)	TTH (days)	Yield (kg)
Data Count	300	300	300	300	300	300
Mean	9.35	2369.81	9.39	2319.66	9.4	2337.86
Std	5.51	1525.63	4.81	1120.44	4.37	961.48
Min	0	200	0	202	0.85	474.98
25%	5	1181.46	6.21	1459.98	6.34	1561.84
50%	9	2262.59	9.39	2225.51	9.33	2268.53
75%	13	3377.69	12.64	3090.02	12.29	2995.75
Max	30	7601.94	26.39	7236.06	22.25	5338.76

Table 5. Result prediction of farmers' field data (harvest time and yield)

Code	Field Data			Result of Model Prediction		TTH	Yield
	TTH (Day)	Area (m ²)	Yield (kg)	TTH (Day)	Yield (kg)	Error Dev.	Error Dev.
F.A (J1)	55	1350	1750	54.6	1292.95	-0.4	-457.05
F.B (J2)	53	875	875	52.8	877.90	-0.2	2.9
F.C (J3)	53	875	875	54.6	997.98	1.6	122.98
F.D (Sis1)	52	875	900	55.5	752.53	3.5	-147.47
F.E (Sis2)	55	875	875	53.8	808.16	-1.2	-66.84
F.F (Suh1)	53	1750	1200	54.8	1274.63	1.8	74.63
F.G (Suh2)	55	1750	1200	53.9	1171.22	-1.1	-28.78
F.H (Suk1)	55	875	875	51.3	767.45	-3.7	-107.55
F.I (Suk2)	55	1350	1350	53.6	1100.47	-1.5	-249.53
F.J (Suk3)	52	875	875	53.5	808.16	1.5	-66.84
F.K (IR1)	55	1350	1500	52.6	1173.69	-2.4	-326.31
F.L (IR2)	55	875	900	51.1	903.77	-3.9	3.77
F.M (R1)	51	1350	1600	53.6	1004.55	2.6	-595.45
F.N (R2)	51	875	1000	52.6	912.92	1.6	-87.08
Average						1.9	166.9

5. CONCLUSION

This study utilized machine learning models, including classical linear regression, random forest, decision trees, and FNN, to estimate shallot harvest time and production yields. Among these models, the linear model demonstrated the highest R-squared value for both estimating harvest time (TTH) and yields. While RidgeCV outperformed the linear model in predicting TTH and yields by a small margin, the ElasticNet model was identified as the best prediction model through the best estimator application. These models effectively estimated the critical factors influencing shallot harvest time and yield. Harvest time is influenced by factors such as plant age and morning temperature, while yield predictions are determined by area, number of beds, bed width, temperature range, bulb weight, and phosphorus weight. Additionally, tacit knowledge such as farmers' calculations for harvest timing, day length, reasons for planting, and confidence in crop yields, was found to significantly impact the model's accuracy. Farmers can leverage this tacit knowledge, including historical data and environmental cues, to estimate the optimal harvest date and predict yield outcomes more accurately.

The study's findings have practical implications for farmers and agricultural planners. By applying these machine learning models and integrating tacit knowledge, farmers can make more informed decisions regarding planting and harvesting schedules, optimizing yield predictions. This could lead to better resource allocation and improved shallot production efficiency. The model performed well in Brebes Regency but may require further adjustments and validation for use in other

regions or with different crops. The forecasted TTH deviation was 23.5%, with a yield inaccuracy of 13%. Comparing these results with actual farm data will be crucial for refining the prediction models and enhancing their applicability in broader agricultural contexts.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to the National Research and Innovation Agency (BRIN), Telkom University, and the Brebes Regency Government for their funding assistance and close collaboration in developing shallot in Brebes and Indonesia. This research would not have been possible without their support. We hope that the results of this collaboration in the form of scientific writing can be useful in supporting government policy-making in developing shallots in Brebes and Indonesia.

REFERENCES

- [1] Utari, M.H., Azijah, Z. (2019). Volatilitas Harga Bawang Merah di Indonesia. *Buletin Ilmiah Litbang Perdagangan*, 13(2): 309-336. <https://doi.org/10.30908/bilp.v13i2.419>
- [2] Konfo, T.R.C., Djouhou, F.M.C., Hounhouigan, M.H., Dahouenon-Ahoussi, E., Avlessi, F., Sohounhloue, C.K.D. (2023). Recent advances in the use of digital technologies in agri-food processing: A short review.

- Applied Food Research, 3: 100329. <https://doi.org/10.1016/j.afres.2023.100329>
- [3] Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A., Uwitonze, C., Ngabonziza, J., Uwamahoro, A. (2023). Crop yield prediction using machine learning models: Case of Irish potato and maize. *Agriculture*, 13(1): 225. <https://doi.org/10.3390/agriculture13010225>
 - [4] Qi, M., Zhang, G. P. (2008). Trend time-series modeling and forecasting with neural networks. *IEEE Transactions on Neural Networks*, 19(5): 808-816. <https://doi.org/10.1109/TNN.2007.912308>
 - [5] Jacques, A.B., Adamchuk, V.I., Cloutier, G., Clark, J.J., Miller, C. (2018). Development of a machine vision yield monitor for shallot onion harvesters. In *Proceedings of the 14th International Conference on Precision Agriculture* June 24–June 27, 2018 Montreal, Quebec, Canada.
 - [6] Selvi, A. (2021). Onion yield prediction based on machine learning. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(2): 2322-2327.
 - [7] Iqbal, L.B., Rahman, M.M., Mamun, S., Nabi, N., Ahamed, M.S. (2022). OnionBangla: A supervised machine learning approach for predicting onion yield using Bangladeshi climate data. In *2022 32nd International Conference on Computer Theory and Applications (ICCTA)*, pp. 110-115. <https://doi.org/10.1109/ICCTA58027.2022.10206199>
 - [8] Lestari, N.A.P., Dijaya, R., Azizah, N.L. (2021). Identification growth quality of red onion during planting period using support vector machine. *Journal of Physics: Conference Series*, 1764(1): 012060. <https://doi.org/10.1088/1742-6596/1764/1/012060>
 - [9] Leukel, J., Zimpel, T., Stumpe, C. (2023). Machine learning technology for early prediction of grain yield at the field scale: A systematic review. *Computers and Electronics in Agriculture*, 207: 107721. <https://doi.org/10.1016/j.compag.2023.107721>
 - [10] Filippi, P., Jones, E.J., Wimalathunge, N.S., Somarathna, P.D., Pozza, L.E., Ugabje, S.U., Jephcott, T.G., Paterson, S.E., Whelan, B.M., Bishop, T.F. (2019). An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agriculture*, 20: 1015-1029. <https://doi.org/10.1007/s11119-018-09628-4>
 - [11] Basso, B., Liu, L. (2019). Seasonal crop yield forecast: Methods, applications, and accuracies. *Advances in Agronomy*, 154: 201-255. <https://doi.org/10.1016/bs.agron.2018.11.002>
 - [12] Van Klompenburg, T., Kassahun, A., Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177: 105709. <https://doi.org/10.1016/j.compag.2020.105709>
 - [13] Sumberg, J., Okali, C., Reece, D. (2003). Agricultural research in the face of diversity, local knowledge and the participation imperative: Theoretical considerations. *Agricultural Systems*, 76(2): 739-753. [https://doi.org/10.1016/S0308-521X\(02\)00153-1](https://doi.org/10.1016/S0308-521X(02)00153-1)
 - [14] Oliveira, R.A.D., Brunetto, G., Loss, A., Gatiboni, L.C., Kürtz, C., Júnior, V.M., Lovato, P.E., Oliveira, B.S., Souza, M., Comin, J.J. (2016). Cover crops effects on soil chemical properties and onion yield. *Revista Brasileira de Ciência do Solo*, 40: e0150099. <https://doi.org/10.1590/18069657rbcs20150099>
 - [15] Dewangan, U., Talwekar, R.H., Bera, S. (2022). Systematic literature review on crop yield prediction using machine & deep learning algorithm. In *2022 5th International Conference on Advances in Science and Technology (ICAST)*, pp. 654-661. <https://doi.org/10.1109/ICAST55766.2022.10039620>
 - [16] Murti, A.C., Al Machfudz, W.D.P., Prihatiningrum, A.E., Arifin, S. (2022). Effect of planting distance and bulb size on growth and production of shallots (*Allium ascalonicum* L.). In *IOP Conference Series: Earth and Environmental Science*, 1104(1): 012002. <https://doi.org/10.1088/1755-1315/1104/1/012002>
 - [17] Hidayah, B.N., Sugianti, T., Mardiana, M., Pramudia, A. (2023). The impact of weather anomalies on shallot seed production in West Lombok, Indonesia. In *E3S Web of Conferences*, 373: 03003. <https://doi.org/10.1051/e3sconf/202337303003>
 - [18] Sumarni, N., Hidayat, A. (2005). *Budidaya Bawang Merah (in Bahasa) Panduan Teknis No 3. vol. 1*. Jakarta: Pusat Penelitian dan Pengembangan Hortikultura, Badan Penelitian dan Pengembangan Pertanian.
 - [19] Mermoud, A., Tamini, T.D., Yacouba, H. (2005). Impacts of different irrigation schedules on the water balance components of an onion crop in a semi-arid zone. *Agricultural Water Management*, 77(1-3): 282-295. <https://doi.org/10.1016/j.agwat.2004.09.033>
 - [20] Patel, N., Rajput, T.B.S. (2013). Effect of deficit irrigation on crop growth, yield and quality of onion in subsurface drip irrigation. *International Journal of Plant Production*, 7(3): 417-436.
 - [21] Fauziah, R., Susila, A.D., Sulistyono, E. (2017). Budidaya Bawang Merah (*Allium ascalonicum* L.) pada Lahan Kering Menggunakan Irigasi Sprinkler pada Berbagai Volume dan Frekuensi. *Jurnal Hortikultura Indonesia*, 7: 1-8. <https://doi.org/10.29244/jhi.7.1.1-8>
 - [22] Cho, Y.C., Lee, J.T., Park, Y.G., Jeong, B.R. (2011). Effect of mulching material and planting density on growth and bulb development of shallot (*Allium cepa* var. *ascalonicum* Backer). *Korean Journal of Plant Resources*, 24(5): 507-513. <https://doi.org/10.7732/kjpr.2011.24.5.507>
 - [23] Sipahutar, T., Hidayat, S., Girsang, M.A., Haloho, L., et al. (2022). Characteristics and analysis of shallots farming in Dolok Silau Simalungun, North Sumatra. *Agric*, 34(2): 287-299. <https://doi.org/10.24246/agric.2022.v34.i2.p287-299>
 - [24] Ayu, N.G., Rauf, A., Samudin, S. (2016). Pertumbuhan dan hasil dua varietas bawang merah (*Allium ascalonicum* L.) pada berbagai jarak tanam. *AGROTEKBIS: JURNAL ILMU PERTANIAN (e-Journal)*, 4(5): 530-536.
 - [25] Harsela, C.N. (2023). Growth and yields of bima brebes shallot variety planted using a floating hydroponics system. *Eduvest-Journal of Universal Studies*, 3(7): 1381-1388. <https://doi.org/10.59188/eduvest.v3i7.887>
 - [26] Hasanah, L.N., Fatah, L., Bachri, A.A., Susanti, H. (2023). SERI methods to measure the vulnerability of shallot farming to climate change in Tapin Regency, South Kalimantan. *Technium Sustainability*, 3: 26-35. <https://doi.org/10.47577/sustainability.v3i.8544>
 - [27] Darnhofer, I., Bellon, S., Dedieu, B., Milestad, R. (2010). Adaptiveness to enhance the sustainability of farming

- systems. A review. *Agronomy for Sustainable Development*, 30: 545-555. <https://doi.org/10.1051/agro/2009053>
- [28] Wandal, A.K., Retang, E.U.K., Saragih, E.C. (2023). Pengaruh Kompetensi dan Motivasi Petani Terhadap Keberhasilan Usahatani Bawang Merah di Kelurahan Maulumbi. In *Proceeding Sustainable Agricultural Technology Innovation (SATI)*, Sumba: Universitas Kristen Wira Wacana Sumba, pp. 168-175.
- [29] Rahayu, E., Irianto, H., Sutrisno, J. (2023). Production and price risk analysis of shallot (*Allium stipitatum* regel) cultivation among farm households in brebes district, Indonesia. *Applied Ecology & Environmental Research*, 21(3): 26252640. https://doi.org/10.15666/aeer/2103_26252640
- [30] Adadi, A. (2021). A survey on data - efficient algorithms in big data era. *Journal of Big Data*, 8(1): 24. <https://doi.org/10.1186/s40537-021-00419-9>
- [31] Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2: 1-21. <https://doi.org/10.1186/s40537-014-0007-7>
- [32] Bonhomme, R. (2000). Bases and limits to using 'degree. day' units. *European Journal of Agronomy*, 13(1): 1-10. [https://doi.org/10.1016/S1161-0301\(00\)00058-7](https://doi.org/10.1016/S1161-0301(00)00058-7)
- [33] Salari, H., Antil, R.S., Saharawat, Y.S. (2021). Responses of onion growth and yield to different planting dates and land management practices. *Agronomy Research*, 1914.
- [34] Asseng, S., Foster, I.A.N., Turner, N.C. (2011). The impact of temperature variability on wheat yields. *Global Change Biology*, 17(2): 997-1012. <https://doi.org/10.1111/j.1365-2486.2010.02262.x>
- [35] Sutardi, Pramono, J., Widodo, S., Martini, T., Alifia, A.D., Apriyana, Y., et al. (2022). Double production of shallot (*Allium cepa* L var. *aggregatum*) based on climate, water, and soil management in sandy land. *International Journal on Advanced Science, Engineering and Information Technology*, 12: 1756. <https://doi.org/10.18517/ijaseit.12.5.14698>
- [36] Apriyanti, D.H., Spreeuwes, L.J., Lucas, P.J. (2023). Deep neural networks for explainable feature extraction in orchid identification. *Applied Intelligence*, 53(21): 26270-26285. <https://doi.org/10.1007/s10489-023-04880-2>

NOMENCLATURE

cat.	Categorical
m	Meter
mo	Microorganism