TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 23, No. 6, December 2025, pp. 1579~1589
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v23i6.27103

Optimization of principal component analysis and k-nearest neighbors in cultivation area classification red onion

Arif Ridho Lubis[1], Purwa Hasan Putra[1], Fahdi Saidi Lubis[2]
[1]Department of Computer Engineering and Informatics, Politeknik Negeri Medan, Medan, Indonesia
[2]Departement of Computer Science and Information Technology, Internasional Islamic University Malaysia, Selangor, Malaysia

Article Info

Article history:

Received Mar 29, 2025
Revised Oct 12, 2025
Accepted Oct 19, 2025

Keywords:

K-nearest neighbor
Optimization
Principal component analysis
Red onion
Region classification

ABSTRACT

This research aims to increase the effectiveness in classifying shallot cultivation areas through the combined application of principal component analysis (PCA) and k-nearest neighbors (KNN) methods. Shallot is an important agricultural commodity, and identification of optimal areas for its cultivation is essential to support food self-sufficiency. Onion cultivation is generally done in the highlands. One of the areas with shallot cultivation in North Sumatra Province is Berastagi, Karo Regency. This research was conducted by determining the spatial extent of upland land. In the use of data there are 2 types of data that will be used: land suitability dataset and land condition dataset for each region. The PCA method is utilized to simplify the data structure by reducing the number of dimensions and removing insignificant attributes, while KNN was used to classify regions based on their suitability for shallot cultivation. This research produces a classification map that can be used to identify the most optimal areas for shallot cultivation. The test results with the regional spatial dataset using precision, recall and fi-score testing accuracy value 0.92%, and macro avg value 0.94%, weighted avg value 0.93%.

Corresponding Author:

Arif Ridho Lubis
Departement of Computer Engineering and Informatics, Politeknik Negeri Medan
Medan, Indonesia
Email: arifridho@polmed.ac.id

1. INTRODUCTION

Red onion (*Allium ascalonicum L.*) is one of the important horticultural commodities in Indonesia, especially in North Sumatra. This plant has high economic value and a wide range of functional benefits, both as a kitchen spice and as a traditional medicine [1], [2]. In addition, shallots also contain bioactive compounds that are beneficial to health, such as antioxidants, antibacterial agents, and antiradicals. Shallots are one of the country's leading commodities with promising market prospects and play a role in supporting food self-sufficiency programs. However, shallot cultivation is greatly influenced by environmental factors, particularly water availability, soil type, and climatic conditions. This plant is quite sensitive to both excess and deficiency of water, which can cause bulb rot or shrinkage [3], [4]. In North Sumatra, areas such as Berastagi in Karo Regency are known as major centers for shallot cultivation because they have highland geographical characteristics that are suitable for the growth of this plant.

Despite their great potential, shallot farmers in Indonesia often face various technical obstacles, including crop failure and significant price fluctuations. For example, the price of shallots rose from Rp 36,320 in January 2024 to Rp 48,400 in May 2024, indicating high market volatility [4], [5]. One of the main causes of crop instability is the lack of information regarding the suitability of land for proper shallot cultivation.

Classifying regions based on land suitability remains a major challenge in Indonesia's agricultural system. Factors such as soil type, altitude, climate, and water availability are critical to the success of cultivation. Therefore, a more scientific and systematic approach is needed to help classify cultivation regions effectively and efficiently [6].

Shallots are widely grown in both lowlands and highlands. Those grown at high altitudes usually have smaller bulbs and less shiny color, and those grown at low altitudes are usually older. One of the most common technical constraints is the suitability of the region in which to grow shallots. Through this process, a classification of shallot cultivation areas will be conducted. This research will implement soil condition data and soil condition with several parameters using machine learning [7], [8].

The classification of regions based on land suitability remains a major challenge in Indonesia's agricultural system. Factors such as soil type, altitude, climate, and water availability are critical to the success of cultivation. Therefore, a more scientific and systematic approach is needed to help classify cultivation areas effectively and efficiently [9], [10]. The combination of these two methods has been proven effective in various previous studies, such as in the classification of onion species and apple varieties with accuracies of 95% and 91.67%, respectively [11], [12]. This study identified 51,499 high-quality variants and utilized the data to build a genomic estimated breeding value (GEBV) model, and applied machine learning methods to predict tuber weight. Validation results on 260 new individuals showed that the model was able to achieve 83.2% prediction accuracy. As an effort to overcome these problems, we also implemented a digital breeding approach based on genomic data from 98 superior onion strains [5].

Another study was conducted with the aim of simplifying the data and removing less relevant attributes without reducing the essence of information from the original data. This was done through the application of the principal component analysis (PCA) method to improve the accuracy performance of the k-nearest neighbor (KNN) classification algorithm. The results of the modified method showed an average accuracy of 88%, with K values varying from $k = 3$ to $k = 9$ [13]. The main contribution of this research is the development of a spatial classification system for shallot cultivation areas based on machine learning, utilizing land condition data and environmental parameters. This research aims to produce a more accurate and applicable classification model as a basis for decision making in planning shallot cultivation in the highlands of North Sumatra.

2. METHOD

In carrying out the research process, a design is needed that can determine the stages in the process of starting the design, data collection, and the process of applying the classification of shallot cultivation areas. The research design as a road map of research that will be completed, a mature design will produce a study that can be useful for the community, especially in this study supporting the government in increasing shallot cultivation and useful for Berastagi farmers in determining the spatial highland land that has an impact on crop yields.

2.1. Datasets

In this study the authors will collect data located in several areas of North Sumatra including in Nagalingga Village, Merek District, Kabanjahe District, Karo Regency, North Sumatra 22173, and in the Lake Toba area. In using the data there are 2 types of data that will be used: land suitability dataset and land condition dataset for each region. Each table, generated from the suitability of features with optimal land planting conditions. For example, a temperature of 25-32 °C, and being at an altitude of 0-450 meters above sea level, as well as other categories, will generally fall into the excellent/good category.

2.2. Preprocessing

Data cleaning is an important step in data pre-processing that aims to ensure the quality of data before it is used in additional analysis or machine learning models. Data cleaning uses various techniques to find and correct errors, inconsistencies, and inaccuracies present in the data set. Data preprocessing is the act of refining and evaluating old data sets to produce new data suitable for use in subsequent processes. The reduced, transformed, and transformed data included in this 70% will be used for testing and 30% for training. To standardize the input dataset features, the training and testing data will be scaled throughout the procedure. To standardize the data, normalized Z scores will be used [14], [15]. This guarantees a reliable comparison of features with different scales but with the same version.

This dataset, which has previously been made available to around 500 users, has five label categories: excellent, very good, fair, poor and very poor. This indicates the extent of the condition of the onion cultivation field. The purpose of this data is to train and test the PCA and KNN models, and will be

used as training data, the features used to obtain the above labels are the condition of the area, and the criteria of the agricultural area.

a. Temperature (Celsius)
b. Rainfall (mm/year)
c. Elevation (above sea level)
d. Drainage (very good, good, fair, bad, very bad)
e. Soil texture (very fine, fine, medium, coarse, very coarse)
f. Potential of hydrogen (pH)
g. Cation exchange capacity (very good, good, fair, poor, very poor)
h. Base saturation (very good, good, fair, poor, very poor)
i. Relief/slope (flat, gentle, sloping, steep, steep)

The research design is shown in Figure 1. In this study, the author used a dataset of land conditions for shallot cultivation based on several criteria, shown in Table 1.

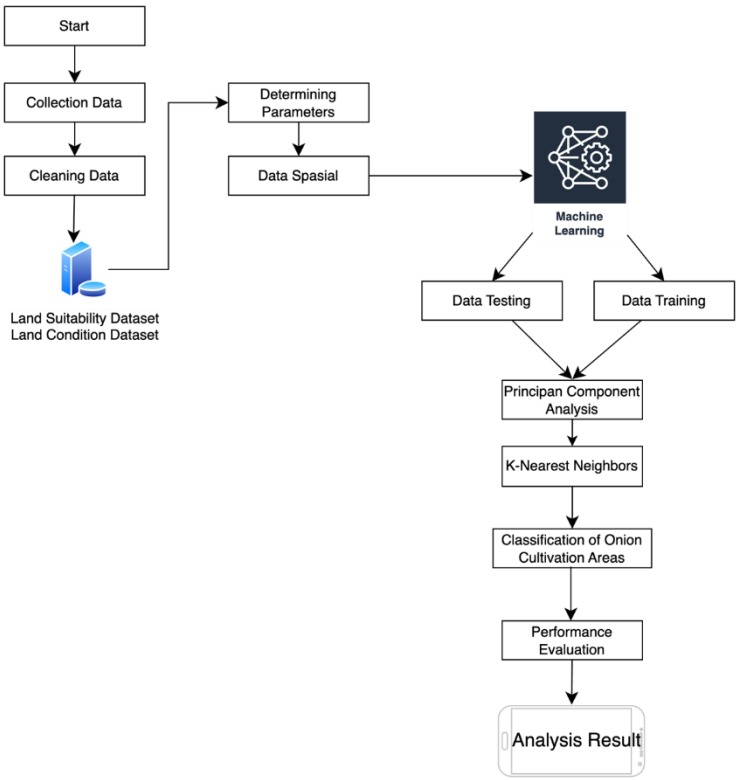

Figure 1. Research process flow model

Table 1. Land condition dataset

Id	Temperature	Rainfall	Elevation	Drainage	Soil texture	pH	Cation exchange capacity	Relief	Label
1	25	1700	400	Very good	Medium	6.1	Very good	Flat	Very good
2	33	2700	1700	Simply	Rough	6.7	Very bad	Italics	Bad
3	28	1300	1000	Both	Medium	5.5	Simply	Ramps	Both
4	30	1100	200	Very good	Smooth	5.9	Very good	Ramps	Very good
…	….	…	…	…	…	…	…	…	…

2.3. Principal component analysis

The dimensionality reduction process is carried out through the application of PCA with the aim of reducing the number of features or variables in the dataset without eliminating the essential information contained therein [16], [17]. The purpose of this process is to speed up processing time, simplify model complexity, and reduce the risk of overfitting. Feature reduction is done by transforming high-dimensional data into lower-dimensional data consisting of uncorrelated attributes [18]. This research proposes a PCA-based framework to select a subset of significant and mutually uncorrelated features. This study utilizes PCA

as a feature selection method to reduce data by retaining features that are independent of each other [19] . A variance calculation is performed to identify the degree of dispersion in the medical dataset, using (1) to determine how much the data deviates in the analyzed sample.

$$Var(x) = \sigma^n = \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{Z}_{ij} - \mu_j\right)^2 \quad (1)$$

Next, the covariance was calculated to determine the relationship between each class. A covariance value close to zero indicates that there is no relationship between the two dimensions. The covariance calculation process is done using (2).

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_{xj})(y_{ij} - \mu_{yj}) \quad (2)$$

The final step involves calculating the eigenvalues and eigenvectors of the covariance matrix. Next, the eigenvalues are transformed through varimax orthogonal rotation using (3).

$$Det\ (A - \lambda I) = 0 \quad (3)$$

This research applies the PCA method to the features contained in the training data and test data.

2.4. K-nearest neighbors

The KNN classification algorithm is utilized to predict the category of data. This process is done by calculating the distance between the test data and the input data to determine the number of nearest neighbors (k). The final category of the data is determined based on the majority of votes obtained from these nearest neighbors [20]. The KNN classifier utilizes a distance metric to calculate the closeness between test and training data. When the number of samples used is smaller, the KNN algorithm is able to produce higher accuracy with a lighter computational burden [21], [22]. The concept of KNN is illustrated in the following diagram, shown in the following Figure 2.

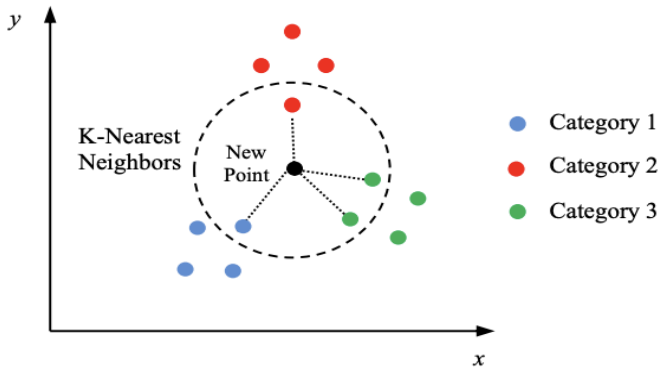

Figure 2. Explanation of KNN

The implementation steps of the KNNs algorithm are:

a. Determine the K value
Determine the value of K, which is the number of nearest neighbors used in the classification process. If the K value is too small, the model tends to be very sensitive to noise in the data, which can lead to overfitting. Conversely, if the K value is too large, the model may lose flexibility and result in underfitting.

b. Calculate the distance between points
For each test data, calculate its distance to all the data in the training dataset. Use a distance metric such as, Euclidean distance (most common):

$$d\ (p,q) = \sqrt{\sum(pi - qi)^2} \quad (4)$$

c. Nearest neighbor identification
   Sort the distance calculation results from smallest to largest. Select the *K* nearest data as the nearest neighbor.
d. Grade or grade prediction
   For classification: take the majority class of the *K* neighbors (using majority voting). For regression: calculate the average value of the *K* neighbors.

2.5. Evaluation

In order to assess how well a classification model works, it is important to use some measure of judgment. Performance evaluation of classification models requires the use of certain relevant metrics. These steps can be determined using the confusion matrix as an evaluation tool. Through analyzing the elements in the confusion matrix, the performance of the classification model can be thoroughly evaluated. In this research, a number of metrics are used as evaluation indicators, including accuracy, precision, recall, F1-score, and area under the curve-receiver operating characteristic (AUC-ROC) [23], [24].

In multiclass classification, accuracy is one of the most commonly used evaluation metrics and is obtained directly from the confusion matrix. Classification models with higher accuracy scores perform better. The accuracy score is calculated based on the number of correct predictions for each category. Accuracy estimation can be obtained through the (5) [25].

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

In dealing with onion datasets, relying on overall accuracy as the sole indicator of classification performance is often inappropriate, especially for detecting positive classes that are in the minority. Therefore, this study uses the precision metric, which is widely applied in data analysis and statistical testing. The precision value is calculated using (6), precision is measured by dividing the number of positive samples correctly classified by the model by the total positive predictions generated. This approach considers the special challenges arising from the presence of minority classes, thus providing a more comprehensive evaluation of the classification performance in the context of the onion dataset.

$$Prec = \frac{TP}{TP+FP} \quad (6)$$

True positive rate, also known as recall in the context of information retrieval, describes the proportion of relevant objects that are correctly recognized out of all objects predicted to be relevant. This metric evaluates the extent to which the classification model is able to accurately identify and retrieve relevant examples from the available dataset. The recall calculation can be formulated as shown in (7).

$$Rec = TP_{rate} = \frac{TP}{TP+FN} \quad (7)$$

In general, F1-score is the harmonic mean of recall and precision values. It provides a balanced measure by considering the model's ability to find relevant data (recall) as well as the accuracy of its predictions (precision). By integrating the two metrics, F1-score provides a more comprehensive evaluation of the classification model's performance, as shown in (8).

$$F1 - Score = \frac{2}{precision^{-1}+recall^{-1}} = 2 \text{ x} \frac{precision \text{ x } recall}{precision+recall} \quad (8)$$

3. RESULTS AND DISCUSSION

This section presents the results of the research accompanied by an in-depth discussion. Results can be displayed through various visual forms such as images, graphs, tables, and other media to facilitate reader understanding. The discussion can be divided into several subsections according to the focus of the analysis.

3.1. Training model

This research was designed by applying the waterfall model, whose illustration is presented in the Figure 3.

Figure 3. Spatial innovation building model

Figure 3 presents the development model of spatial innovation of shallot cultivation starting from the analysis stage, which includes identifying the needs of innovative products based on the application of spatial control, both in the form of hardware and software. The next stage is innovation design, which includes the design of system inputs, processes, and outputs. Next, the implementation stage is carried out through the coding process, followed by the testing stage to evaluate the functionality and success of system access. After testing, the innovation is socialized and put into operation. The final stage of this model is the system refinement process, which is carried out over five months of implementation.

Data is obtained from actual data for each district/city from the internet. There are some data features that are rarely available, such as drainage, pH, cation exchange capacity, and base saturation. The rest are available on the internet. For the rarely available data above, an estimate of the average land condition in the province of North Sumatra was made (e.g the average pH is around 5.6 - 6.5, so it was chosen within that range). The following is a dataset of land conditions for each region indicated in Table 2.

Table 2. Land condition dataset for each region

Id	Regional	Temperature	Rainfall	Elevation	Drainage	Soil texture	pH	Cation exchange capacity	Base saturation	Relief
1	Karo	24	1200	1000	Very good	Smooth	5.6	Very good	Very good	Ramps
2	Simalungun	27	1600	1500	Very good	Smooth	5.8	Both	Both	Ramps
3	Samosir	23	1400	1600	Both	Medium	6.3	Very good	Very good	Flat
4	Tapanuli Utara	25	1200	1200	Very good	Medium	5.8	Very good	Very good	Italics
5	Sibolga	26	1400	200	Very good	Medium	6.4	Very good	Very good	Italics
…	….	…	…	…	…	…	…	…	…	

At the following stage is before the data is trained, where the dataset has been managed and is ready for training such as the following data. The following training data before training is shown in Table 3.

Table 3. Regional spatial data

Id	Regional	Temperature	Rainfall	Elevation	Drainage	Soil texture	pH	Cation exchange capacity	Base saturation	Relief
1	Karo	24	1200	1000	Very good	Smooth	5.6	Very good	Very good	Ramps
2	Simalungun	27	1600	1500	Very good	Smooth	5.8	Both	Both	Ramps
3	Samosir	23	1400	1600	Both	Medium	6.3	Very good	Very good	Flat
4	Tapanuli Utara	25	1200	1200	Very good	Medium	5.8	Very good	Very good	Italics
5	Sibolga	26	1400	200	Very good	Medium	6.4	Very good	Very good	Italics
…	….	…	…	…	…	…	…	…	…	

Figure 4 shows a visualization of the land suitability map per region. To display the map, a shapefile containing the latitude and longitude of each boundary line on the map was used, then plotted using matplotlib. The district colors and numbers correspond to the regional dataset that has been processed and predicted by the trained KNN model. The image is a visualization of the results of dimensionality reduction using PCA, which projects the data into two main components: PCA component 1 on the horizontal axis and PCA component 2 on the vertical axis. The colored dots represent the training data used to train the model,

with different colors indicating different classes or labels. Meanwhile, the cross (X) indicates the test data, where the red X represents the model's prediction results, while the other colored Xs represent the original labels of the test data. From the distribution of the points, it can be seen that data with similar colors tend to group (form clusters), indicating similar characteristics between data within the same class. Most of the red Xs are close to a specific color cluster, indicating that the model is able to predict fairly well. However, there are also red Xs that are far from the appropriate cluster, indicating a prediction error. Visualizations like these are useful for understanding the fit of the test data distribution to the training data and for identifying areas where the model performs optimally and areas that are prone to errors.

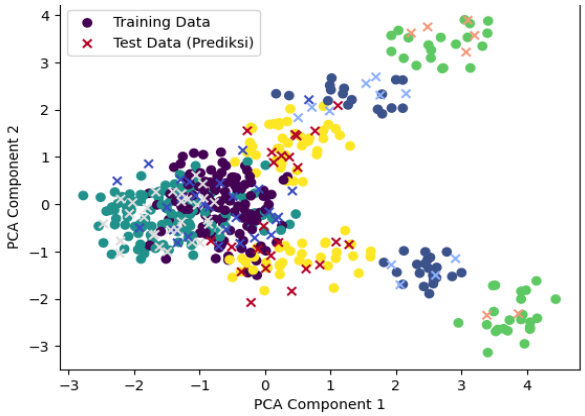

Figure 4. Spatial implementation results with trained data

3.2. Testing results

After the data training process is carried out using a predetermined method, the next stage is to test each model. This test aims to evaluate the performance of the model in classifying and determining potential areas for shallot cultivation accurately. The results of this stage will be the basis for assessing the effectiveness of the approach used in this research. The following is the result data after the training is shown in Table 4.

Table 4. Results of spatial testing of regional data

Id	Regional	Temperature	Rainfall	Elevation	Drainage	Soil texture	pH	Cation exchange capacity	Base saturation	Relief	Label
1	Karo	24	1200	1000	1.0	0.0	5.6	1.0	1.0	Ramps	4
2	Simalungun	27	1600	1500	1.0	0.0	5.8	0.0	0.0	Ramps	4
3	Samosir	23	1400	1600	0.0	2.0	6.3	1.0	1.0	Flat	4
4	Tapanuli Utara	25	1200	1200	1.0	2.0	5.8	1.0	1.0	Italics	4
5	Sibolga	26	1400	200	1.0	2.0	6.4	1.0	1.0	Italics	0
…	….	…	…	…	…	…	…	…	…		

The predicted labels (0.12) will be decoded into categories (good and bad). Then each category will be assigned a color. Figure 5 shows the distribution of land suitability levels in North Sumatra Province based on the categories "moderate," "good," and "very good." Yellow represents areas with moderate land suitability, light green indicates good land suitability, and dark green indicates very good land suitability. Each region/city is numbered according to the legend, such as Asahan, Batu Bara, Deli Serdang, Mandailing Natal, and others. This map shows that areas with very good land suitability are spread across several districts, such as Mandailing Natal, Padang Lawas, and parts of North Labuhan Batu. Conversely, areas with moderate suitability are more widespread in the northern and western parts of North Sumatra. This map is useful for land use planning, agriculture, and regional development, as it can help determine areas with optimal potential for specific activities based on their land suitability levels. Next are the model performance results for each test with the regional spatial dataset shown in Table 5.

Table 5. Model performance results for each test with regional spatial dataset

Dataset: classification of onion cultivation areas				
Authors	Precision	Recall	F1-score	Support
0	0.79	0.92	0.85	24
1	1.00	0.92	0.96	12
2	0.93	0.85	0.89	33
3	1.00	1.00	1.00	7
4	1.00	1.00	1.00	24
Accuracy			0.92	100
Macro Avg	0.94	0.94	0.94	100
Weighted Avg	0.93	0.92	0.94	100

Overall, the model achieved 92% accuracy. The macro average for precision, recall, and F1-score was 0.94. High F1-score values across almost all classes indicate good model performance and balance between classes, although there was a slight decrease in classes 0 and 2, likely due to the slightly lower number of correct predictions compared to other classes.

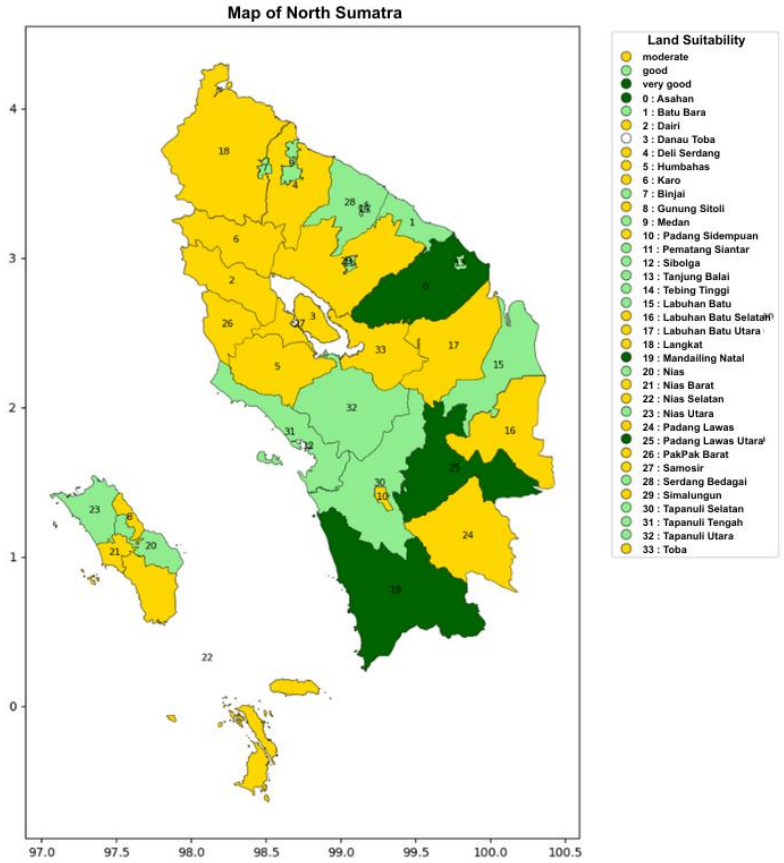

Figure 5. Red onion land classification results

Based on the evaluation results shown in the shallot cultivation area classification graph, it can be seen that the precision, recall, and F1-score values in the model tend to be at a low and relatively flat range. This indicates that the model's ability to identify areas suitable for shallot cultivation is still not optimal. The model is not yet able to distinguish well between truly suitable and unsuitable areas, so the level of precision and the level of success of the model in finding all true positive data (recall) are both still low. The low F1-score value indicates that the balance between precision and recall has not been achieved, so the model is not stable in its performance shown in Figure 6. Based on the test results with the regional spatial dataset using precision, recall and F-score testing, the accuracy value is 0.92%, and the macro avg value is 0.94%, the weighted avg value is 0.93%.

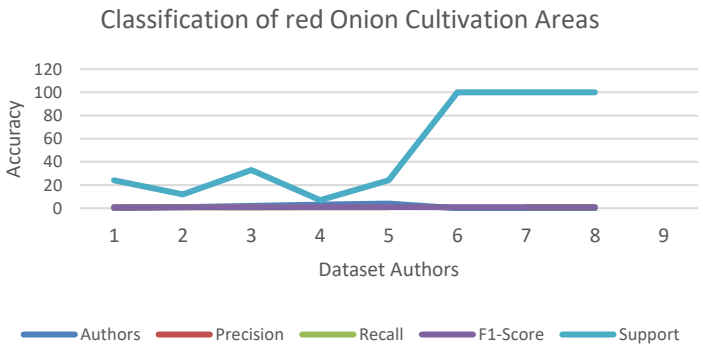

Figure 6. Model evaluation results

4. CONCLUSION

This study aims to address the problems that exist in shallot cultivation, it is necessary to pay attention to the suitability and characteristics of land in shallot cultivation. Onion planting is generally done in the highlands. One of the areas with shallot cultivation in North Sumatra Province is Berastagi, Karo Regency. This research was conducted by determining the spatial extent of upland land. In the use of data there are 2 types of data that will be used: land suitability dataset and land condition dataset for each region. Each table is generated by matching features with optimal land cultivation conditions. For example, a temperature of 25-32 °C, and an altitude of 0-450 meters above sea level. The combination of PCA and KNN methods is proven to be able to increase efficiency and accuracy in the process of classifying shallot cultivation areas. This research produces a classification map that can be used to identify the most optimal areas for shallot cultivation. By identifying optimal areas, this research contributes to efforts to increase shallot production and achieve food self-sufficiency. From the research results it can be seen, where spatial locations that have dark green color codes are very good areas for shallot cultivation based on modeling and specified variables. The results of testing with regional spatial datasets using precision, recall and fi-score testing accuracy value 0.92%, and macro avg value 0.94%, weighted avg value 0.93%.

ACKNOWLEDGMENTS

The authors express their appreciation for the funding support provided through DIPA funds of Politeknik Negeri Medan for the fiscal year 2024.

FUNDING INFORMATION

The authors express their appreciation for the funding support provided through DIPA funds of Politeknik Negeri Medan for the fiscal year 2024, in accordance with the contract number: B/425/PL5/PT.01.05/2024, Along with all team members who play a role in the implementation of collaborative research through this PTPI scheme.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Arif Ridho Lubis	✓	✓			✓	✓			✓	✓				✓
Purwa Hasan Putra		✓	✓						✓		✓			
Fahdi Saidi Lubis						✓				✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The author declares no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

Research related to human use has complied with all relevant national regulations and institutional policies in accordance with the principles of the Declaration of Helsinki and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, [Arif Ridho Lubis]. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

REFERENCES

[1] N. A. Basallote *et al*., "IoT-based growth analysis of red onion in controlled hydroponic environment using electroculture," in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, IEEE, Nov. 2022, pp. 1–6, doi: 10.1109/TENCON55691.2022.9977614.

[2] M. Y. Puchkov, M. M. Shagapov, and K. K. Eshajieva, "Technologies of cultivation of red onion varieties," *BIO Web of Conferences*, vol. 71, p. 1006, Nov. 2023, doi: 10.1051/bioconf/20237101006.

[3] M. Dutta, D. Gupta, S. Juneja, A. Nauman, and G. Muhammad, "Comparative growth analysis of onion in deep water culture and soil based systems: Enhancing medicinal plant cultivation in Urbanized environments," *IEEE Access*, vol. 12, pp. 38202–38218, 2024, doi: 10.1109/ACCESS.2024.3373787.

[4] M. M. Hasan, M. T. Zahara, M. M. Sykot, R. Hafiz, and M. Saifuzzaman, "Solving onion market instability by forecasting onion price using machine learning approach," *2020 International Conference on Computational Performance Evaluation, ComPE 2020*, pp. 777–780, 2020, doi: 10.1109/ComPE49325.2020.9200033.

[5] J. Choi *et al.*, "Genotype-driven phenotype prediction in onion breeding: machine learning models for enhanced bulb weight selection," *Agriculture (Switzerland)*, vol. 14, no. 12, p. 2239, Dec. 2024, doi: 10.3390/agriculture14122239.

[6] A. A. Waoo and M. Tiwari, "Optimizing crop yield and fertility management using KNN and Ml," *ShodhKosh: Journal of Visual and Performing Arts*, vol. 5, no. 5, pp. 380–385, May 2024, doi: 10.29121/shodhkosh.v5.i5.2024.1888.

[7] S. L. Bairwa *et al*., "Economic analysis of Onion production in Bihar: Insight into cost of cultivation, post-harvest management and constraint identification," *International Journal of Agricultural Invention*, vol. 9, no. 1, pp. 273–277, Jun. 2024, doi: 10.46492/ijai/2024.9.1.34.

[8] N. Tziolas, S. A. Ordoudi, A. Tavlaridis, K. Karyotis, G. Zalidis, and I. Mourtzinos, "Rapid assessment of anthocyanins content of onion waste through visible-near-short-wave and mid-infrared spectroscopy combined with machine learning techniques," *Sustainability (Switzerland)*, vol. 13, no. 12, p. 6588, Jun. 2021, doi: 10.3390/su13126588.

[9] I. Luthfifah, E. Hernawati, H. Nugroho, B. R. Aditya, and P. Aji, "Implementation of K-nearest neighbors algorithm for red onion crop suitability land assessment in Selaawi, Indonesia," *Ingenierie des Systemes d'Information*, vol. 30, no. 2, pp. 301–306, Feb. 2025, doi: 10.18280/isi.300201.

[10] X. Cheng, M. Liao, and J. Liu, "Geographical origin identification of panax notoginseng using a modified K-nearest neighbors model with near-infrared spectroscopy," *IEEE Access*, vol. 13, pp. 13832–13846, 2025, doi: 10.1109/ACCESS.2025.3530766.

[11] D. N. T. Ronquillo, J. M. S. Selda, and M. V. C. Caya, "Classification of onion species in the philippines using a convolutional neural network (CNN)," in *2023 IEEE 15th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM 2023*, IEEE, Nov. 2023, pp. 1–6, doi: 10.1109/HNICEM60674.2023.10589121.

[12] M. A. Hasan, "Classification of apple types using principal component analysis and K-nearest neighbor," *International Journal of Information System Technology and Data Science*, vol. 1, no. 1, pp. 15–22, Jun. 2023, doi: 10.61398/ijist-das.v1i1.11.

[13] A. H. Lubis, P. Sihombing, and E. B. Nababan, "Analysis of accuracy improvement in K-Nearest neighbor using principal component analysis (PCA)," *Journal of Physics: Conference Series*, vol. 1566, no. 1, p. 12062, Jun. 2020, doi: 10.1088/1742-6596/1566/1/012062.

[14] A. A. Abu-Shareha, H. Qutaishat, and A. Al-Khayat, "A framework for diabetes detection using machine learning and data preprocessing," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1654–1667, Dec. 2024, doi: 10.47738/jads.v5i4.363.

[15] A. Brijith, "Data preprocessing for machine learning," *CSIM*, vol. 3, 2023.

[16] R. H. Maharrani, P. D. Abda'u, and M. N. Faiz, "Clustering method for criminal crime acts using K-means and principal component analysis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 34, no. 1, pp. 224–232, Apr. 2024, doi: 10.11591/ijeecs.v34.i1.pp224-232.

[17] D. Selvida, A. F. Pulungan, and A. S. Huzaifah, "Optimization of garlic cultivation land selection using PCA and K-means approach in spatial intelligent system," *Eastern-European Journal of Enterprise Technologies*, vol. 3, no. 2 (135), pp. 54–64, June 2025, doi: 10.15587/1729-4061.2025.325340.

[18] M. M. Arcinas *et al*., "An efficient course recommendation system for higher education students using machine learning techniques," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 14, no. 2, pp. 1468–1475, Apr. 2025, doi: 10.11591/eei.v14i2.7711.

[19] T. Huang, J. Li, and W. Zhang, "Application of principal component analysis and logistic regression model in lupus nephritis patients with clinical hypothyroidism," *BMC Medical Research Methodology*, vol. 20, no. 1, p. 99, Dec. 2020, doi: 10.1186/s12874-020-00989-x.
[20] P. H. Putra, M. S. Novelan, and M. Rizki, "Analysis K-nearest neighbor method in classification of vegetable quality based on color," *Journal of Applied Engineering and Technological Science*, vol. 3, no. 2, pp. 126–132, 2022, doi: 10.37385/jaets.v3i2.763.
[21] M. S. Novelan, S. Efendi, P. Sihombing, and H. Mawengkang, "Vehicle routing problem optimization with machine learning in imbalanced classification vehicle route data," *Eastern-European Journal of Enterprise Technologies*, vol. 5, no. 3 (125), pp. 49–56, Oct. 2023, doi: 10.15587/1729-4061.2023.288280.
[22] L. N. Lidya, A. M. P. Arif, and A. Y. Dini, "A new triple-weighted k-nearest neighbor algorithm for tomato maturity classification," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 4, pp. 955–962, Aug. 2025, doi: 10.29207/resti.v9i4.6441.
[23] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.05756.
[24] I. Markoulidakis and G. Markoulidakis, "Probabilistic confusion matrix: a novel method for machine learning algorithm generalized performance analysis," *Technologies*, vol. 12, no. 7, p. 113, July 2024, doi: 10.3390/technologies12070113.
[25] S. Sathyanarayanan, "Confusion matrix-based performance evaluation metrics," *African Journal of Biomedical Research*, pp. 4023–4031, Nov. 2024, doi: 10.53555/ajbr.v27i4s.4345.

BIOGRAPHIES OF AUTHORS

Arif Ridho Lubis received a Doctoral degree from Universitas Sumatra Utara in 2023, a Masters degree from Universiti Utara Malaysia in 2012, and graduated from Universiti Utara Malaysia in 2011, both in Information Technology. He is a lecturer at the Department of Computer Engineering and Informatics, Politeknik Negeri Medan in 2015. His research interests include computer science, networking, and project management science. He can be contacted at email: arifridho@polmed.ac.id.

Purwa Hasan Putra was born in Medan, Indonesia, in 1991. He is a lecturer at the Computer Engineering Study Program of Politeknik Negeri Medan. He earned his Master's degree from the University of North Sumatra in 2019. His main research interests are data mining, machine learning, deep learning. He can be contacted at email: pputra@polmed.ac.id.

Fahdi Saidi Lubis received a Bachelor's degree in Information Technology from Universitas Sumatera Utara in 2021 and a Master's degree in Information System from Telkom University in 2023. He is currently pursuing a Doctor of Philosophy (Ph.D.) in Computer Science and Information Technology at the International Islamic University Malaysia (IIUM). His research interests include data governance, ICT governance, and learning analytics. He can be contacted at email: fahdilubis@gmail.com.