Spatial Classification of Sentinel-2 Satellite Images with Machine Learning Approach

by Achmad Fauzan

Submission date: 12-Apr-2023 07:19AM (UTC+0700)
Submission ID: 2062030420
File name: Geoplanning_Dea.docx (694.89K)
Word count: 5936
Character count: 32720

GeoPlanning
Journal of Geomatics and Planning
Vol. x, No. x, year

e-ISSN: 2355-6544

Received: .....;
Accepted: .....;
Published: .....

Keywords:
Satellite Imagery, Classification, Machine Learning
*Corresponding author(s)
email: achmadfauzan@uii.ac.id

Original Research open access

Spatial Classification of Sentinel-2 Satellite Images with Machine Learning Approach

Dea Ratu Nursidah [1], Achmad Fauzan [1*], Marcelinus Alfafisurya Setya Adhiwibawa [2]

1. *Statistics Department, Faculty of Mathematics and Natural Science, Universitas Islam Indonesia, Indonesia.*
2. *Agricultural Data System Scientist, PCTC, Mondelez International*

DOI: 10.14710/geoplanning.8.1.pp-pp

Abstract

This study aims to classify buildings and non-buildings from Sentinel-2 Satellite Images using a Machine Learning approach. The limitations of the machine learning method for classification used in this study are the Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT) methods. The three methods' results are compared to find the best method in the classification process. Furthermore, the proportion between buildings and non-buildings around Universitas Islam Indonesia was calculated from the best method's results. The results are in the form of a classification with four indicators, namely level of accuracy, sensitivity, specificity, and Area Under the Curve (AUC). We found that the best performing method in this study is the SVM method based on the average accuracy results, the smallest average variance difference in the variance of training and testing data, and three other indicators from the number of iterations accomplished. In the density proportion, we concluded that the closer the distance to UII campus, the greater the percentage of buildings. As for non-buildings, the farther from the center point, the higher the rate of non-buildings.

1. Introduction

Buildings are physical forms resulting from construction work that are integrated with their domicile, part, or all of which are above ground and/or water which function as places for humans to carry out their activities, whether for residence or dwelling, religious activities, business activities, social activities, culture or specific activities (Pemerintah Republik Indonesia, 2002). In every regional development endeavor, the establishment of educational institutions is one of the attractions for people in the area and from outside the area. Consequently, residential areas must be built. The Special Region of Yogyakarta is home to many tertiary institutions, resulting in a high level of urbanization. Universitas Islam Indonesia (UII) is one of the universities located in Yogyakarta. The establishment of UII's integrated campus since 1993 has presented an academic community that requires supporting services such as rent, boarding houses, restaurants, photocopying, shopping centers, and other facilities. It is this demand that opens opportunities for the community to improve the economy hence many buildings were erected around the UII area.

In constructing buildings, it is necessary to observe areas that have the potential to become targets for service support. One of the ways is to use spatial data. Spatial data is data about the spatial aspects of an object, phenomenon, or event on the earth's surface but has yet to have a specific reference or coordinate system (Ramdani, 2021). Spatial data has an overview of the area on the earth's surface and is represented in digital format in the form of rasters and vectors with specific values (Supuwiningsih et al., 2022). Raster data is data generated from a remote sensing system where geographical objects are represented as a grid cell structure

called a pixel (Irwansyah, 2013). In contrast, according to Awangga (Awangga, 2017), vector data is stored in point coordinates, which displays, locates, and stores spatial data using points, lines, or polygons. One application of spatial data is the measurement of building density from satellite imagery in certain areas. The spatial data used in this study is Sentinel-2 satellite imagery obtained from the European Space Agency (ESA). Sentinel-2 has 13 bands with their characteristics. Four bands have 10m spatial resolution, six bands have 20m spatial resolution, and three bands have 60m resolution. Spatial resolution is the smallest size that can still be detected by an imaging system (Danoedoro, 2012). Then spatial data be analyzed using a data mining algorithm, classification. The accuracy value is an essential point in determining the algorithm's accuracy.

Some of the related studies include Antara et.al (Antara et al., 2022), conducted research on the classification of rice fields and non-rice fields using the random forest method on Sentinel-1 SAR imagery with a spatial resolution of 20 meters (the sensing system can distinguish objects that are 20 meters or more apart), the results of the cross validation showed that the classification results have the same accuracy for both rice and non-rice fields, with a percentage value of 96.90%. Firmansyah et al (2019) conducted a study comparing the Support Vector Machine and Decision Tree methods to map mangroves using Sentinel-2B Satellite Imagery. Sampurno (Sampurno & Thoriq, 2016) conducted a classification analysis using Landsat 8 Operational Land Imager (OLI) image data using the Maximum Likelihood Classification (MLC) method. The land cover classes are built-up land, rice fields before harvest, newly planted rice fields, shrubs, dense forest cover, medium forest cover, mixed forest, mixed gardens, open land, and water bodies. Awaliyan(Awaliyan & Sulistioadi, 2018) used sentinel-2A with the tree algorithm method. The algorithm parameter used in determining class separation is the Normalized Difference Vegetation Index (NDVI).

Compared to previous research, in this study, classification of buildings and non-buildings was carried out using Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT) classification methods. This study uses SAR sentinel two images with a resolution of 10x10 so that the object used will be clearer, using two objects, namely buildings and non-building. From the results of the three methods, a comparison was made by looking at the average accuracy and the smallest average variance difference in the from training and testing data, and three other indicators from the number of iterations accomplished (recall/ sensitivity, specificity, and Area Under the Curve (AUC)). Furthermore, a buffer analysis was carried out to find out the density proportion between buildings and non-buildings based on predetermined radius and points.

2. Data and Methods

2.1. Sentinel-2

Sentinel-2 is a satellite launched through a collaboration between The European Commission and the European Space Agency in the Global Monitoring for Environment and Security (GMES) program. To be able to offer information on the most recent conditions of the Earth and space for environmental and security applications, this satellite was launched to monitor the state of the Earth's surface. Landsat 5/7, SPOT-5, SPOT-Vegetation, and Envisat MERIS missions, whose operational lives are about to expire, will be continued by Sentinel-2. The mission is to provide high spatial and temporal resolution satellite imagery so that users can still obtain the latest Earth surface scanning data (Semedi et al., 2021).

On Sentinel-2, 13 bands are installed with their respective characteristics. Four bands having a spatial resolution of 10m ensure compatibility with SPOT 4/5 and meet the usability requirements for land coverage classification. Six bands have a spatial resolution of 20m which is a requirement for other level 2 processing parameters. The 60m resolution band is used exclusively for atmospheric correction and cloud screening (443nm for aerosols, 940nm for water vapor, and 1375nm for detecting wispy clouds). A spatial resolution of 60m is considered sufficient to capture the spatial variability of atmospheric geophysical parameters (Oktaviani & Kusuma, 2017).

2.2. *Support Vector Machine, Logistic Regression, and Decision Tree*

The Support Vector Machine (SVM) is a classification method whose objective is to identify a hyperplane that may most effectively divide two classes. The margin value should be maximized to produce a decent hyperplane. The margin is the separation between the support vector and the hyperplane. Linear classifier is the main principle in support vector machines (Suyanto, 2017). The first step of an SVM algorithm is to define the equation of a separating hyperplane which is Equation 1.

$$w.x_i + b = 0 \quad (1)$$

where w: weight vector, x_i: i-th data, b: bias value. If the b value is used as an additional weight for w_0, the following formula is used.

$$w_0 + w.x_i = 0 \quad (2)$$

Every point located above or below the hyperplane used the Equation 3 and 4 (Han et al., 2012).

$$w_0 + w.x_i > 0 \quad (3)$$

$$w_0 + w.x_i < 0 \quad (4)$$

The weights will be adjusted so that the hyperplane will separate into 2 classes. y_i is the i-th data class, so the Equation 5 and 6 is used.

$$H_1: w_0 + w.x_i \geq 1 \; for \; = +1 \quad (5)$$

$$H_2: w_0 + w.x_i \geq -1 \; for \; y_i = -1 \quad (6)$$

The largest margin can be obtained by maximizing the distance between the hyperplane and its closest point $\frac{1}{\|w\|}$. This can be formulated as a Quadratic Programming (QP) problem, which is to find the minimum point of Equation 7 considering the constraints of Equation 8.

$$\min_w \tau(w) = \frac{1}{2}\|w\|^2 \quad (7)$$

$$y_i(x_i.w + b) - 1 \geq 0 \quad (8)$$

This equation can be solved with the Langrage Multiplier using Equation 9.

$$L(w,b,a) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l}(a_i((x_i.w.b) - 1)) \quad (9)$$

a_i is a Langrage Multiplier that has a value of 0 or a positive value $a_i \geq 0$. The optimal value can be calculated by minimizing L with respect to w and b. and maximize L over a_i. By using L=0 to describe the properties at the optimal point, Equation 9 can be modified by maximizing the problem contained in a_i, using Equation 10.

$$\sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j x_i x_j \quad (10)$$

where $a_i \geq 0 (i = 1,2,\ldots,l)$ $\sum_{i=1}^{l} a_i y_i = 0$. From the results of this calculation, we can obtain a_i, which is mostly positive. Data that is correlated with a positive a_i is what is referred to as a support vector (Nugroho et al., 2003).

Logistic Regression (LR) is a data analysis method used to find the relationship between the dependent variable (y) which has a binary category and the independent variable (x) which is polychotomous. The output of the dependent variable consists of 2 categories which are usually denoted by y = 1, representing a success and y = 0, representing failure (Hosmer & Lemeshow, 2000). The probability function for every observation is presented in Equation 11.

$$f(y) = \begin{cases} \pi^y(1-\pi)^{1-y} & \text{for } y = 0,1 \\ 0 & \text{for } y \neq 0,1 \end{cases} \quad (11)$$

π = Probability of success. If $y = 0$ then $f(y) = (1-\pi)$, and if $y = 1$ then $f(y) = \pi$. The probability fuction for logistic regression can be defined as Equation 12.

$$f(z) = \begin{cases} \dfrac{1}{1+e^{-z}} & \text{for } -\infty < z < +\infty \\ 0 & \text{others} \end{cases} \quad (12)$$

$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$. In general, logistic regression models are written in Equation 13.

$$\pi(x) = \begin{cases} \dfrac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}} & \text{for } 0 \le \pi(x) \le 1 \\ 0 & \text{for } \pi(x) \le 0 \text{ and } \pi(x) > 1 \end{cases} \quad (13)$$

where π(x) is the probability of a successful event with $0 \le \pi(x) \le 1$, under the condition that 1 is declared a successful event and 0 means a failure. $\pi(x)$ is a nonlinear function, so it needs to be transformed into logit form. Parameter estimation of the logistic regression model can be described using the logit transformation of $\pi(x)$.

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \quad (14)$$

$$\frac{\pi(x)}{1-\pi(x)} = e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)} \quad (15)$$

$$g(x) = logit\{\pi(x)] = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (16)$$

g(x) is the relationship function of the LR model which is referred to as the logit relationship function. The expected probability of $Y = 1$ (success) based on the given value of X is.

$$\pi_i = P(Y_i 1|X) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}} \quad (17)$$

β_0 is a constant and β_1 is the coefficient from each variable. The Odds Ratio (OR) is given by Equation 18.

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k} \quad (18)$$

OR > 1 is more likely to be regarded as $Y = 1$ (success), OR < 1 is more likely to be regarded as $Y = 0$ (failure), OR = 0.5 indicates that relationships between Y and X are nonexistent (Utari, 2019).

The Decision Tree (DT) classifies data items into a limited number of predetermined classes (Werdiningsih et al., 2020). The DT C5.0 algorithm is an improvement on the Decision Tree C4.5, using a reduced tree to make the actions taken more succinct. The new decision tree technique, ID3 and C4.5, is superior in terms of memory savings and pruning (Muflikah et al., 2021). The C5.0 algorithm outperforms C4.5 in terms of speed, memory, and effectiveness. A classification technique suitable for big data sets is the C5.0 algorithm. In the C5.0 algorithm, attribute selection is processed using the gain ratio. The gain ratio measure is used to select test attributes at each node in the tree. The attribute with the highest gain ratio will be selected as the parent for the next node. The steps for creating a tree in the C5.0 algorithm are like creating a tree in the C4.5 algorithm. The similarity includes entropy and gain calculations. If the algorithm stops at the gain calculation, then the C5.0 algorithm will continue by calculating the gain ratio using the existing gain and entropy (Pratiwi et al., 2020). Equation 19 is the formula used to calculate entropy value:

$$\text{Entropy (S)} = \sum_{i=1}^{n} -pi * log2pi \quad (19)$$

S: Set of cases, n: Number of partitions S, pi: Si to S ratio. To obtain the gain value, the following formula is used:

$$Gain(\text{S}, A) = \text{Enthropy } (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * \text{Entropy}(Si) \quad (20)$$

A: Attribute being used, $|Si|$: Number of cases at the i-th partition, $|S|$: Number of cases in S. The following is the formula used to calculate the gain ratio (Harani & Damayanti, 2021).

$$\text{Gain Ratio} = \frac{\text{Gain}\ (S, A)}{\sum_{i=1}^{n} \text{Entropy}(Si)} \quad (21)$$

Gain (S, A): gain value from variable, $\sum_{i=1}^{n} Entropy(Si)$: Number of entropy values in a particular variable.

2.3. Model Evaluation

Evaluation model using holdout method. The holdout method divides the data set randomly into two independent (non-overlapping) subsets, the division of which is usually 2/3 of the training data and the remaining 1/3 of the test data. However, different portions can also be used according to certain considerations (Suyanto, 2017). From the results of training and testing, the confusion matrix is used. The Confusion Matrix (CM) provides comparative information between the classification results from the model output and the actual classification results. The four components that represent the classification results in a CM are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) (Syahrani et al., 2019).

Table 1. Confusion Matrix

Confusion Matrix		Actual Values	
		True	False
Predicted Values	True	True Positive (TP)/ Correct result	False Positive (FP)/ Unexpected result
	False	False Negative (FN)/ Missing result	True Negative (TN)/ Correct absence of result

Evaluation of the goodness of the model used is measured using four criteria, namely: values of accuracy, recall/sensitivity, specificity, and Area Under the Curve (AUC) from the construction of the classification model.

Accuracy is the value of the accuracy of predictions whose model is correct. The following is a calculation to find the accuracy value. The higher the accuracy value, the better the classification (Widayati et al., 2021). A recall is a ratio predicted to be correct or relevant when compared to all correct data. The higher the sensitivity value, the less likely the results of the positive class classification are wrong (Zhu et al., 2010). Specificity is the correctness of predicting a negative compared to the overall negative data. The higher the specificity value, the better the classification performance for predicting because it has low false positives (Maxim et al., 2014).

The FPR is the percentage of negative cases in the data incorrectly reported as positive (the likelihood of a false warning appearing). The total number of negative cases that were incorrectly reported as positive cases was divided by the total number of negative cases. By calculating the likelihood of an output from a randomly chosen sample of the positive population, AUC assesses discriminatory performance. AUC values will always range from 0 to 1. The greater the AUC value, the stronger the classification (Defiyanti & Jajuli, 2015). calculation of each evaluation size, presented in the Equation 22 to 26.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (22)$$

$$\text{Recall/Sensitivity} = \frac{TP}{TP + FN} \times 100\% \quad (23)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100\% \quad (24)$$

$$\text{FPR} = 1 - \text{Specificity} \quad (25)$$

$$\text{AUC} = \frac{1 + \text{sensitivity} - FPR}{2} \quad (26)$$

2.4. Research Methodology

The data used in this research is Sentinel-2 satellite image data on July 7, 2021, with an area directly adjacent to the UII, sourced from https://scihub.copernicus.eu/. The Sentinel-2 Satellite Image used is an image

in which there are not many clouds so that it can be used for research. The population used is satellite imagery in area directly adjacent to UII, and the sample used is the area directly adjacent to UII in 2021. Five variables are used in this study: variable x as Universal Transverse Mercator (UTM) easting/ longitude, variable y as UTM northing/latitude, independent variable (X), namely band 8 and band 11, and dependent variable (Y), namely class with 0 as non-building and 1 as a building. Band 8 has a Visible and Near Infrared (VNIR) wavelength with a spatial resolution of 10 m, while Band 11 has a Short-Wave Infrared (SWIR) wavelength with a spatial resolution of 20 m (Fletcher & European Space Agency., n.d.). These two bands are used because both are constituents of the Normalized Difference Built-Up Index (NDBI), and NDBI extraction has been carried out in previous studies (Chen et al.; Deng & Wu, 2012; Guo et al., 2015; Hidayati et al., 2018) The research flow chart is presented in the Figure 1.

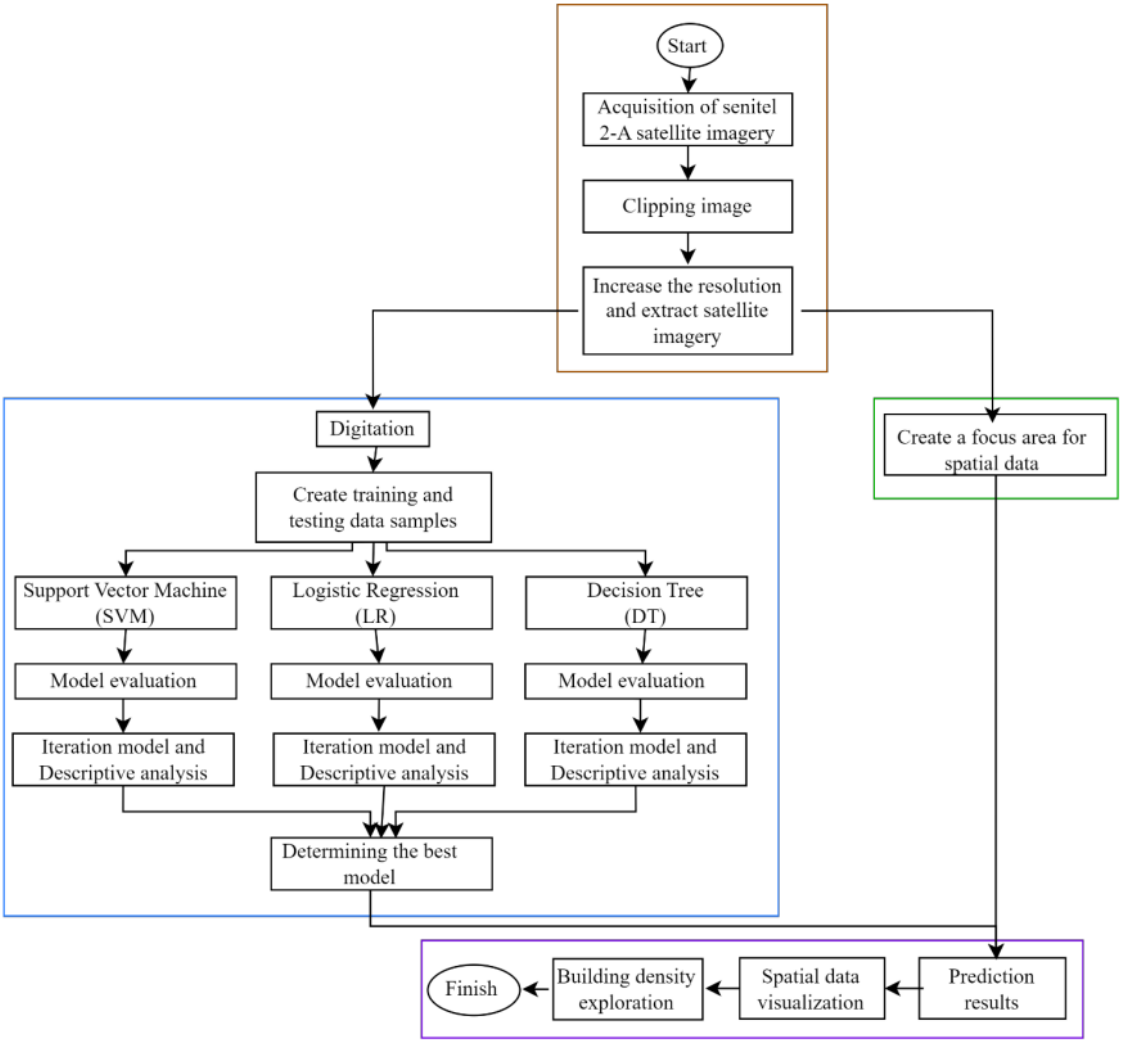

Figure 1. Research flowchart

The brown color is the first stage, namely preprocessing. At the preprocessing stage, geoprocessing is carried out. Geoprocessing is a process in Geographic Information System (GIS) that is used to process analysis of spatial data, which will eventually produce novel data and information. Vector geoprocessing are

geoprocessing techniques implemented on vector data structures. Examples of vector geoprocessing techniques include clipping, buffering, splitting, merging, and overlaying vectors (Marjuki, 2014). Moreover, the following steps are carried out in the first stage. (1) Retrieval of data through the Copernicus web. (2) Image cropping using the SNAP application. The image is cropped according to the area to be used. (3) Improved image resolution on band 11 and extracted band 8 using the SNAP application. The resolution increase in band 11 is done because band 11 has a resolution of 20 m while band 8 has a resolution of 10 m, so to make the two bands have the same resolution, an increase will be made in band 11.

In the green box, data is generated that borders the UII. In making this complete data, a merger between bands 8 and band 11 was then carried out by clipping or cutting using a polygon area directly adjacent to the village of Umbulmartani. Next, the following steps are carried out in the blue box stage. (1) Digitize buildings and non-buildings by creating polygons in QGIS. Digitization is the process of converting an analog map into a digital map using a digitization table. The way it works is by converting the existing spatial features on the map into a set of x, y coordinates. To produce accurate data, high-quality analog map sources are needed. For the digitization process, high precision and concentration are required from the operator (Adil, 2017). (2) Making samples of training and testing data by conducting random points in areas that have been digitized. (3) Classification analysis using the support vector machine, logistic regression, and decision tree methods with 40 iterations. Iteration here is used to see the pattern and stability of the predictions obtained. (4) Descriptive analysis using the iteration results of each classification. (5) Comparing methods using the average indicator of each method to obtain the best method.

While in the last stage (purple box), the following steps are carried out. (1) Predict the best methods and datasets to obtain predictive images with raster visualization. (2) Performing buffering analysis to calculate the proportion of buildings and non-buildings. Buffering is the creation of zones or areas using the distance from an object. Buffering involves the ability to create zones using the distance from a selected object, such as a point, line, or area (polygon). Buffers are polygon shapes because they represent the area around an object. Buffering also refers to the formation of sona or corridors in raster data models (Prahasta, 2002).

Software used for classification most use R software because of its several advantages, such as convenience, portability, multiplatform, and programmability (Rosadi, 2016; Sa'adah et al., 2021). The stages of spatial data visualization and preprocessing are used by QGIS, because QGIS is an open source that provides many functional capabilities and features, and the number continues to grow, including supporting various raster and vector formats (Bruy & Svidzinska, 2015; Flenniken et al., 2020; Zarodi & Anshori, 2018).

3. Result and Discussion

3.1. Data Preparation

Area clipping is done to obtain the area of study, which is directly adjacent to the Umbulmartani village. We merged regional bands directly adjacent to UII as the campus is included in the village of Umbulmartani, the areas directly adjacent to the village of Umbulmartani are Sardonoharjo, Sukoharjo, Widomartani, Harjobinangun, Pakembinangun, and Wukirsari villages. The raw data has 524,521 points with 4 variables: longitude, latitude, band 8/ B8 (VNIR), and band 11/ B11 (SWIR). There was a merger of regional bands directly adjacent to UII because the campus is included in the village of Umbulmartani, the areas directly adjacent to the village of Umbulmartani are Sardonoharjo, Sukoharjo, Widomartani, Harjobinangun, Pakembinangun, and Wukirsari villages.

The preparation of training and testing data is carried out by taking random points from the digitized results. The total data was 2000 which consisted of 1000 buildings and 1000 non-buildings. Figure 2 shows the results of digitization for buildings and non-buildings. The green polygons represent buildings, and the brown polygons represent non-buildings. Digitization is carried out to obtain training data and testing data which are

labeled 0 for non-buildings and 1 for buildings. We split the data into 80 percent training data with 1600 rows and 20 percent testing data with 400 rows.

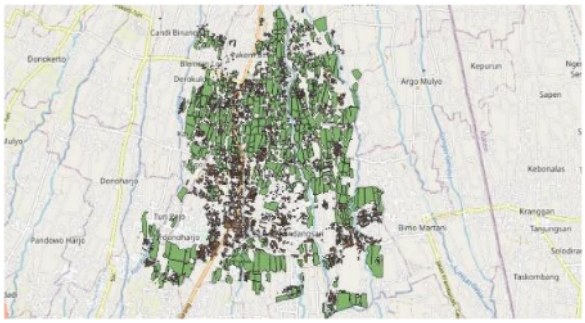

Figure 2. Digitization

3.2. Classification

After the data is divided into training and testing data, it is continued with the three methods of classification. Because each processing produces a different output value (because it contains random numbers), 40 iterations are carried out to see the pattern or stability of each method. Following are the resulting evaluation values of each iteration and classification. Table 4 shows values of accuracy, sensitivity, specificity, and AUC of each iteration in the SVM method classification.

Table 4. Evaluation results of Support Vector Machine model

Iteration	Training data				Test Data			
	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
1	91.75	91.06	92.46	91.80	89.50	89.74	89.27	89.50
2	91.44	90.92	91.97	91.40	90.25	90.26	90.24	90.30
3	91.69	90.96	92.45	91.70	90.00	89.80	90.20	90.00
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
38	90.56	89.88	91.27	90.60	94.00	94.53	93.47	94.00
39	91.00	90.34	91.69	91.00	92.50	92.78	92.23	92.50
40	91.00	90.23	91.78	91.00	92.25	93.07	91.41	92.20

A classification will be carried out in the logistic regression method, based on Equation 14, the resulting form of the logistic regression equation is presented in Equation 27.

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = -1.486 - 0.004X_1 + 55.201\,X_2$$
$$\pi(x) = \frac{e^{-1.486-0.004X_1+55.201\,X_2}}{1+e^{-1.486-0.004X_1+55.201\,X_2}} \quad (27)$$

The calculated odds ratio value is presented in Equation 28.

$$\frac{\pi}{1-\pi} = e^{-1.486-0.004X_1+55.201\,X_2} \quad (28)$$

For example, if the x_1 value is 2496 and x_2 is 0.183, an odds ratio value of 0.183 is obtained, meaning that the data is included in Y(0) (non-building). Table 5 shows values of accuracy, sensitivity, specificity, and AUC of each iteration using the training and testing data. Table 5 the results are presented LR method classification.

Table 5. Evaluation results of LR using Test data (a) and Train data (b)

Iteration	Training				Testing			
	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
1	91.75	91.27	92.25	91.80	89.00	89.23	88.78	89.00
2	91.44	91.02	91.87	91.40	90.25	89.85	90.64	90.20
3	91.50	90.93	92.09	91.50	90.50	89.90	91.09	90.50
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
38	90.62	89.99	91.28	90.60	93.50	94.03	92.96	93.50
39	90.94	90.33	91.57	91.00	92.50	92.78	92.23	92.50
40	91.06	90.55	91.58	91.10	92.00	92.61	91.37	92.00

While Table 6 shows values of accuracy, sensitivity, specificity, and AUC in DT method classifictoion.

Table 6. Evaluation results of Decision Tree using Test data (a) and Train data (b)

Iteration	Training				Testing			
	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
1	92.19	92.06	92.32	92.20	87.75	88.54	87.02	87.80
2	92.56	92.97	92.15	92.60	91.00	89.60	92.42	91.00
3	92.38	92.00	92.27	92.40	89.25	88.83	89.66	89.20
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
38	91.25	90.11	92.45	91.30	91.75	91.67	91.84	91.80
39	91.50	91.54	91.46	91.50	91.75	91.79	91.71	91.80
40	91.06	93.82	88.67	91.20	91.25	95.21	87.74	91.50

3.3. Method Comparison

After obtaining the classification results from each method, we selected the best method based on the average accuracy and the smallest average variance difference in the from training and testing data, and three other indicators from the number of iterations accomplished (sensitivity, specificity, and AUC), which is shown in Table 7.

Table 7. Mean and Variance comparison.

Metrics Value	Mean						Variance					
	SVM		LR		DT		SVM		LR		DT	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Accuracy (%)	91.31	91.24	91.26	91.18	91.96	90.44	0.09	1.37	0.08	1.35	0.20	1.72
Sensitivity	90.80	91.13	90.80	91.06	92.14	91.02	0.13	3.17	0.14	3.28	0.57	5.78
Specificity	91.83	91.37	91.76	91.30	91.78	89.94	2.46	0.08	2.50	1.35	4.15	2.46
AUC	91.31	91.23	91.28	91.18	91.97	90.48	0.10	1.35	0.09	1.37	0.19	1.65

Based on the comparison of the mean values from all classifications we found that the decision tree model has the highest accuracy. However, the SVM model has the smallest difference in mean values between classifications on the train and test data, which is 0.06. Besides that, it has a relatively small difference in variance compared to other methods. Therefore, we declared SVM as the best performing method in this study.

3.4. Predictive Images

Predictive images are used to make predictions with the best method selected, namely the support vector machine into intact data. The illustration prediction results are obtained as Table 8.

Table 8. Six Predicted Data from The Dataset (*an illustration*).

Longitude (UTM)	Latitude (UTM)	Class
437895	9156275	1
437905	9156275	1
437915	9156275	0
437925	9156275	0
437935	9156275	0
437945	9156275	0

Table 8 shows the SVM prediction results for the entire dataset where in the initial dataset there were no class variables because predictions had been made, so the prediction results contained 524,521 data with longitude, latitude, and class variables as the predicted outcome variables. The prediction results obtained will be used as raster data again. The results obtained from the predictions portray nonbuildings in white and buildings in red which is presented in Figure 2.

Figure 2. Prediction Results.

Figure 2 shows the prediction results when zoomed in. Conclusively, the predictions are in accordance with the original base map. After the classification results and prediction results are obtained, one of its uses is to calculate proportions.

3.5. Proportional Calculation

Buffering analysis is performed to identify the relationship between a point and the surrounding area. In this study we drew a buffer with a center point at UII and zones with radii of 1, 2, 3, 4, and 5 km are made. Figure 3 is a figure showing the buffer results.

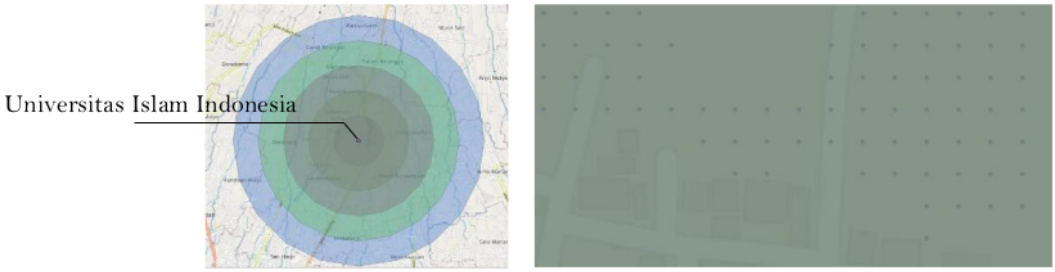

Figure 3. Point spatial buffer from UII.

Buffers are obtained using a distance of 1-5 km where Figure 6 is an example of the points that are in the buffer. The yellow dots are building points while the blue dots are non-building points. The buffering analysis on the complete data is presented in Table 9.

Table 9. Buffering Analysis.

Distance (km)	Area (Km2)	Building Points	Nonbuilding Points	Percentage of Buildings (%)	Percentage of Nonbuildings (%)
1	3092	14735	16159	47.695	52.305
2	12369	46030	69806	39.737	60.263
3	27831	79542	159529	33.271	66.729
4	49477	112767	244121	31.597	68.403
5	77308	136239	309037	30.597	69.403

Based on Table 9, it is found that at a radius of 1 km the percentage of buildings was 47.695% while the remaining 52.305% were nonbuildings. At a radius of 2 km, the percentage of buildings was 39,737% while the remaining 60,263% were nonbuildings. At a radius of 3 km, the percentage for buildings was 33,271% while 66,729% of points were nonbuildings. At a radius of 4 km, the percentage of buildings was 31,597% while 68,403% of points were nonbuildings. At a radius of 5 km, the percentage for buildings was 30.597% while 69.403% of points were nonbuildings.

We arrived at the conclusion that the closer to the central point, the greater the percentage of buildings while for nonbuildings, the farther away, the greater the percentage. This suggests that the UII campus area serves as a new residential center so. Many buildings are erected to support services for the academic community in addition to entrepreneurs who want to create a business that requires premises often target areas close to UII to open their businesses. Areas with dense populations increase community activities which cause traffic congestion, high water demand, large garbage production, high production of water waste, plastic waste, etc. So, it is urgent that the relevant agencies carry out rearrangement and maintenance to reduce the impact that occurs because of the establishment of many buildings.

4. Conclusion

Based on the iteration result, the average values for accuracy, sensitivity, specificity, and AUC in the training data, classified using the SVM method, are 91.3073%, 90.7990%, 91.83%, 91.3125% respectively, while in the testing data each are 91.2438 %, 91.1333%, 90.3683%, 91.2250%. The average values for accuracy, sensitivity, specificity, and AUC in the training data, classified using the LR method, are 91.2630%, 90.7988%, 91.7563%, 91.2750% respectively while in the testing data each are 91.1750%, 91.0563 %, 91.3043%, 91.1825%. The average values for accuracy, sensitivity, specificity, and AUC of the training data, classified using the DT method, are 91.9250%, 92.1365%, 91.7845%, 91.9675, respectively, while the testing data are 90.4375%, 91.0190 respectively. %, 89.9415%, 90.4850%. The best method used in this study is the SVM method which was determined based on the average accuracy and the smallest average variance difference in the from training and testing data, and three other indicators from the number of iterations accomplished (sensitivity, specificity, and AUC). Buffer analysis on the complete dataset showed that at a radius of 1 km, the percentage for buildings was 47.695% while for nonbuildings was 52.305%. At a radius of 2 km, a percentage of 39,737% was obtained for buildings while for nonbuildings it was 60,263%. At a radius of 3 km, the percentage of buildings was 33,271% while nonbuildings were at 66,729%. At a radius of 4 km, the percentage for buildings is 31,597%, while the remaining 68,403% of points were nonbuildings. At a radius of 5 km, the percentage for buildings is 30.597% while the remaining 69.403% were nonbuildings. We concluded for both datasets that the closer to the center point, the greater the percentage of buildings. In contrast, the farther away from the center point, the greater the percentage of nonbuildings.

5. Acknowledgments

We thank all the parties who have provided support and funding for this research.

6. References

Adil, A. (2017). *Geographic Information System*. Yogyakarta: ANDI.

Antara, I. M. O. G., Kusmiyarti, T. B., Suyarto, R., & Wiyanti. (2022). Classification of Rice Field and Non-rice Field using Random ForestMethod on Sentinel-SAR Imagery, Case study in Kediri District, Tabanan Regency, Bali. *Prosiding Seminar Nasional Geomatika VI*, 245–252. Cibinong: Badan Informasi Geospasial RI.

Awaliyan, M. R., & Sulistioadi, Y. B. (2018). Classification of Land Cover in Satellite Imagery Sentinel-2a using The Tree Algorithm Method. *ULIN: Jurnal Hutan Tropis*, *2*(2), 98–104.

Awangga, M. (2017). *Introduction to Geographic Information Systems: Basic Concepts and GIS Builder Applications*. Bandung: Rolly Book series in Geospatial Intellegence.

Bruy, A., & Svidzinska, D. (2015). *QGIS with Example*. Packt.

Chen, X. L., Zhao, H. M., Li, P. X., & Yin, Z. Y. (2006). Remote sensing image-based analysis of the relationship between urban heat island and land use/cover changes. *Remote Sensing of Environment*, *104*(2), 133–146. https://doi.org/10.1016/j.rse.2005.11.016.

Danoedoro, P. (2012). *Pengantar Penginderaan Jauh Digital*. Yogyakarta: ANDI.

Defiyanti, S., & Jajuli, M. (2015). Integration of Classification and Clustering Methods in Data Mining. *Konferensi Nasional Informatika (KNF)*, 39–44.

Deng, C., & Wu, C. (2012). BCI: A biophysical composition index for remote sensing of urban environments. *Remote Sensing of Environment*, *127*, 247–259. https://doi.org/10.1016/j.rse.2012.09.009.

Firmansyah, S., Gaol, J., & Susilo, S. B. (2019). Comparison of SVM and Decision Tree Classifier with Object Based Approach for Mangrove Mapping to Sentinel-2B Data on Gili Sulat, Lombok Timur. *Journal of Natural Resources and Environmental Management*, *9*(3), 746–757. https://doi.org/10.29244/jpsl.9.3.746-757.

Flenniken, J. M., Stuglik, S., & Iannone, B. V. (2020). Quantum GIS (QGIS): An introduction to a free alternative to more costly GIS platforms. *Edis*, *2020*(2), 7. https://doi.org/10.32473/edis-fr428-2020.

Fletcher, Karen., & European Space Agency. (n.d.). *Sentinel-2 : ESA's optical high-resolution mission for GMES operational services*.

Guo, G., Wu, Z., Xiao, R., Chen, Y., Liu, X., & Zhang, X. (2015). Impacts of urban biophysical composition on land surface temperature in urban heat island clusters. *Landscape and Urban Planning*, *135*, 1–10. https://doi.org/10.1016/j.landurbplan.2014.11.007.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Wyman Street, Waltham, MA 02451, USA: Elsevier.

Harani, N. H., & Damayanti, F. S. (2021). Implementation of the C5.0 Algorithm to Determine Potential Customers at the Cimahi Post Office. *Jurnal Sistem Informasi Dan Teknologi*, *4*(1), 69–76. Retrieved from http://www.jurnal.umk.ac.id/sitech.

Hidayati, I. N., Suharyadi, R., & Danoedoro, P. (2018). Combined Image Index for Analysis of Built-up Land and Urban Vegetation. *Majalah Geografi Indonesia*, *32*(1), 24. https://doi.org/10.22146/mgi.31899.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression*. United States of America: Jogn Wiley & Sons, Inc.

Irwansyah, E. (2013). *Geographic Information Systems: Basic Principles and Application Development*. Yogyakarta: Digibooks.

Marjuki, B. (2014). *Geographic Information System Using Quantum GIS 2.0.1 Durfour*. Kementerian Pekerjaan Umum dan Perumahan Rakyat.

Maxim, L. D., Niebo, R., & Utell, M. J. (2014, November 1). Screening tests: A review with examples. *Inhalation Toxicology*, Vol. 26, pp. 811–828. Informa Healthcare. https://doi.org/10.3109/08958378.2014.955932

Muflikah, L., Widodo, Mahmudi, W. F., & Solimun. (2021). *Machine Learning in Bioinformatics*. Malang: Universitas Brawijaya.

Nugroho, A. S., Witarto, A. B., & Handoko, D. (2003). Application of Support Vector Machine in Bioinformatics. *Indonesian Scientific Meeting*. Gifu-Japan. Retrieved from http://asnugroho.net.

Oktaviani, N., & Kusuma, H. A. (2017). Introduction of Sentinel-2 Satellite Imagery for Marine Mapping. *Oseana*, *XLII*(3), 40–55.

Pemerintah Republik Indonesia. *Law of the Republic of Indonesia Number 28 of 2002 concerning Buildings.*, (2002).

Prahasta, E. (2002). *Basic Concepts of Geographic Information Systems*. Bandung: Informatika.

Pratiwi, R., Hayati, M. N., & Prangga, S. (2020). Comparison of C5.0 Algorithm Classification with Classification and Regression Tree (Case Study: Social Data of Family Head of Teluk Baru Village, Muara Ancalong District in 2019). *BAREKNG: Jurnal Ilmu Matematika Dan Terapan*, *14*(2), 267–278. https://doi.org/10.30598/barekengvol14iss2pp267-278.

Ramdani, F. (2021). *Geospatial Data Science: Geospatial Data Processing and Analysis using R*. Malang: Pena Persada.

Rosadi, D. (2016). *Statistical analysis with R* (1st ed.). Yogyakarta: Gadjah Mada University Press.

Sa'adah, U., Rochayati, M. Y., Lestari, D. W., & Lusia, D. A. (2021). *Complete Analysis of Data Mining Algorithms and Their Implementation Using R*. Malang: UB Press.

Sampurno, R. M., & Thoriq, A. (2016). Land Cover Classification using Landsat 8 Operational Land Imager (OLI) Data in Sumedang Regency. *Jurnal Teknotan*, *10*(2), 61–70.

Semedi, B., Rijal, S. S., Sambah, A. B., & Isdianto, A. (2021). *Introduction to Marine Remote Sensing*. Surabaya: Universitas Brawijaya.

Supuwiningsih, N. N., Januhari, N. N. U., Suniantara, I. K. P., & Hanief, S. (2022). *Integration of Spatial Data and Non-Spatial Data of Geographic Information Systems*. Bandung: Media Sains Indonesia.

Suyanto. (2017). *Data Mining for Data Classification and Clustering*. Bandung: Informatika.

Syahrani, I. M., Kusuma, W. A., & Wahyuni, S. (2019). Comparation Analysis of Ensemble Technique With Boosting(XGBOOST) and Bagging(Random Forest) for Classify Splice Junction DNA Sequence Category. *Jurnal Penelitian Pos Dan Informatika*, *9*(1), 27.

Utari, D. T. (2019). *Applied regression analysis with R*. Yogyakarta: Universitas Islam Indonesia.

Werdiningsih, I., Nuqoba, B., & Muhammadun. (2020). *Data Mining Using Android, Weka, and SPSS*. Surabaya: Airlangga University Press.

Widayati, Y. T., Prihati, Y., & Widjaja, S. (2021). Analysis and Comparison of Naïve Bayes and C4.5 Algorithms for MNC Play Customer Loyalty Classification in Semarang City. *TRANSFORMTIKA*, *18*(2), 161–172.

Zarodi, H., & Anshori, M. (2018). *Geographic Information System Training Module using QGIS*. Yogyakarta: SinauGIS.

Zhu, W., Zeng, N., & Wang, N. (2010). Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations. *NESUG Proceedings: Health Care and Life Sciences*. Baltimore, Maryland.

Spatial Classification of Sentinel-2 Satellite Images with Machine Learning Approach

ORIGINALITY REPORT

21%	13%	11%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Bengkulu Student Paper	6%
2	ejournal.kresnamediapublisher.com Internet Source	1%
3	www.ijisrt.com Internet Source	1%
4	Muhammad Haqqi Ghufran Rifaldi, Erwin Budi Setiawan. "Competence Classification of Twitter Users Using Support Vector Machine (SVM) Method", 2019 7th International Conference on Information and Communication Technology (ICoICT), 2019 Publication	1%
5	join.if.uinsgd.ac.id Internet Source	1%
6	Ni Nyoman Supuwiningsih, Adam Sheni Martua Malau Pase. "GEOGRAPHIC INFORMATION SYSTEM MAPPING SENIOR HIGH SCHOOL AND VOCATIONAL HIGH	1%

SCHOOL IN BALI ISLAND BASED INTERNET BY USING GOOGLE MY MAPS", International Journal of Engineering Technologies and Management Research, 2022
Publication

7	vtext.valdosta.edu Internet Source	1 %
8	Tri Wida Amaliya, Putu Artama Wiguna. "Building permit mismatch analysis: A case study of building permits in the city of Surabaya, Indonesia", AIP Publishing, 2022 Publication	1 %
9	www.coursehero.com Internet Source	1 %
10	hal-paris1.archives-ouvertes.fr Internet Source	<1 %
11	www.sba.oakland.edu Internet Source	<1 %
12	Drusch, M.. "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services", Remote Sensing of Environment, 20120515 Publication	<1 %
13	repositorio.iscte-iul.pt Internet Source	<1 %
14	www.science.gov Internet Source	

		<1%
15	Behar, E.. "Screening for generalized anxiety disorder using the Penn State Worry Questionnaire: a receiver operating characteristic analysis", Journal of Behavior Therapy and Experimental Psychiatry, 200303 Publication	<1%
16	Submitted to School of Business and Management ITB Student Paper	<1%
17	media.neliti.com Internet Source	<1%
18	Christina Corbane, Vasileios Syrris, Filip Sabo, Panagiotis Politis et al. "Convolutional neural networks for global human settlements mapping from Sentinel-2 satellite imagery", Neural Computing and Applications, 2020 Publication	<1%
19	Chomsin S. Widodo, Agus Naba, Muhammad M. Mahasin, Yuyun Yueniwati, Terawan A. Putranto, Pangeran I. Patra. "UBNet: Deep learning-based approach for automatic X-ray image detection of pneumonia and COVID-19 patients", Journal of X-Ray Science and Technology, 2021 Publication	<1%

20	Mitali Yeshwant Joshi, Auline Rodler, Marjorie Musy, Sihem Guernouti, Mario Cools, Jacques Teller. "Identifying urban morphological archetypes for microclimate studies using a clustering approach", Building and Environment, 2022 Publication	<1 %
21	Submitted to University of Witwatersrand Student Paper	<1 %
22	Submitted to Institut Pertanian Bogor Student Paper	<1 %
23	Nurhamidah Nurhamidah, Bujang Rusman, Bambang Istijono. "A Raster-based model for flood inundation mapping on delta lowland", MATEC Web of Conferences, 2018 Publication	<1 %
24	researchcommons.waikato.ac.nz Internet Source	<1 %
25	123dok.com Internet Source	<1 %
26	www.fit.vut.cz Internet Source	<1 %
27	www.ijert.org Internet Source	<1 %
28	Submitted to University of Westminster Student Paper	<1 %

29	www.dovepress.com Internet Source	<1%
30	S A Komarudin, D Anggraeni, A Riski, A F Hadi. "Classification of genetic expression in prostate cancer using support vector machine method", Journal of Physics: Conference Series, 2020 Publication	<1%
31	Venn Yan Ishak Ilwaru. "ANALISIS REGRESI LOGISTIK ORDINAL TERHADAP FAKTOR-FAKTOR YANG MEMPENGARUHI WAKTU KELULUSAN MAHASISWA S1 DI FMIPA UNPATTI AMBON TAHUN 2016 DAN 2017", BAREKENG: Jurnal Ilmu Matematika dan Terapan, 2019 Publication	<1%
32	Jiajun Bu, Xin Shen, Bin Xu, Chun Chen, Xiaofei He, Deng Cai. "Improving Collaborative Recommendation via User-Item Subgroups", IEEE Transactions on Knowledge and Data Engineering, 2016 Publication	<1%
33	Mauro Pazmiño Betancourth, Victor Ochoa-Gutiérrez, Heather M. Ferguson, Mario González-Jiménez et al. "Diffuse reflectance spectroscopy for predicting age, species, and insecticide resistance of the malaria mosquito	<1%

	Anopheles gambiae s.l", Research Square Platform LLC, 2023 Publication	
34	dl.icdst.org Internet Source	<1 %
35	jgrs.eng.unila.ac.id Internet Source	<1 %
36	Submitted to Vels University Student Paper	<1 %
37	Yulianto Mustaqim, Ema Utami, Suwanto Raharjo. "Analysis of Daubechies Wavelet and Neural Network for Audio Classification", 2019 International Conference on Information and Communications Technology (ICOIACT), 2019 Publication	<1 %
38	journal.ipb.ac.id Internet Source	<1 %
39	link.springer.com Internet Source	<1 %
40	oseana.lipi.go.id Internet Source	<1 %
41	www.frontiersin.org Internet Source	<1 %
42	www.mdpi.com Internet Source	<1 %

43	www.serena.unina.it Internet Source	<1%
44	A S Thoha, O A Lubis O, D L N Hulu, T Y Sari, Z Mardiyadi. "Utilization of UAV technology for mapping of mangrove ecosystem at Belawan, Medan City, North Sumatera, Indonesia", IOP Conference Series: Earth and Environmental Science, 2022 Publication	<1%
45	Dela Youlina Putri, Rachmadita Andreswari, Muhammad Azani Hasibuan. "Analysis of Students Graduation Target Based on Academic Data Record Using C4.5 Algorithm Case Study: Information Systems Students of Telkom University", 2018 6th International Conference on Cyber and IT Service Management (CITSM), 2018 Publication	<1%
46	Mohanad Hassan Edan, Ruba Muhsen Maarouf, Jabbar Hasson. "Predicting the impacts of land use/land cover change on land surface temperature using remote sensing approach in Al Kut, Iraq", Physics and Chemistry of the Earth, Parts A/B/C, 2021 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		