GeoPlanning
Journal of Geomatics and Planning

Geoplanning: Journal of Geomatics and Planning, Vol. 12, No. 2, 2-25, 253-266

e-ISSN: 2355-6544

Received: 12 April 2023;
Revised: 08 May 2025;
Accepted: 26 October 2025;
Available Online: 31 October 2025;
Published: 31 October 2025.

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.12.2.253-266

Abstract

Urban expansion and land use change are increasingly critical issues in developing regions, where rapid development often leads to unplanned growth and environmental challenges. Accurate and timely classification of built-up and non-built-up areas is essential for supporting sustainable spatial planning and resource management. This study aims to classify built-up and non-built-up areas from Sentinel-2 satellite imagery using a machine learning approach and to analyze their spatial distribution around the Universitas Islam Indonesia (UII) campus. Three machine learning algorithms—Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT)—were applied to perform the classification, and their performances were evaluated using four metrics: accuracy, sensitivity, specificity, and Area Under the Curve (AUC). Among these, the SVM method demonstrated the best performance based on the highest average accuracy, the smallest variance difference between training and testing datasets, and consistent results across multiple iterations. Using the classification results from the best-performing model, a spatial density proportion analysis was conducted. The findings revealed a clear spatial trend: areas closer to the UII campus exhibited a higher proportion of built-up land, while areas located farther from the center had a greater share of non-built-up land. These results confirm the effectiveness of the SVM algorithm for land cover classification using Sentinel-2 imagery and offer valuable insights into urban development patterns in the study area. The outcomes of this research can inform urban planners and policymakers in developing data-driven strategies for sustainable land use, infrastructure development, and campus-centered regional growth planning.

1. Introduction

Built-up areas are physical forms resulting from construction work that are integrated with their domicile, part, or all of which are above ground and/or water that 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 (Law No. 28 of 2002 on Buildings, 2002). In every regional development endeavor, the establishment of educational institutions is one of the attractions for people in and outside the area. Consequently, residential areas must be constructed. The Special Region of Yogyakarta is home to many tertiary institutions, resulting in a high level of urbanization. The Universitas Islam Indonesia (UII) is university located in Yogyakarta. The establishment of the 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. This demand opened opportunities for the community to improve the economy hence many built-up were erected around the UII area.

When constructing built-up areas, it is necessary to observe areas that have the potential to become targets for service support. One such method is to use spatial data. Spatial data are data about the spatial aspects of an object, phenomenon, or event on the Earth's surface, but have yet to have a specific reference or coordinate system (Janipella et al., 2019). Spatial data provide an overview of the area on the earth's surface and are represented in digital format in the form of rasters and vectors with specific values (Pebesma & Bivand, 2023).

Raster data are generated from a remote sensing system where geographical objects are represented by a grid cell structure called a pixel (Bettinger et al., 2017). In contrast, according to Yao & Li (2018), vector data are stored in point coordinates, that display, locate, and store spatial data using points, lines, or polygons. One application of spatial data is the measurement of built-up density using satellite imagery in certain areas. The spatial data used in this study were Sentinel-2 satellite imagery obtained from the European Space Agency (ESA). Sentinel-2 had 13 bands with their characteristics. Four bands have spatial resolution of 10 m, six bands have spatial resolution of 20 m, and three bands had 60m resolution. The spatial resolution is the smallest size that can still be detected by an imaging system (Athanasiou et al., 2017). Then, spatial data were analyzed using a data mining algorithm and classification. The accuracy value is an essential factor in determining the accuracy of on algorithm.

Some of the related studies include Zhai et al. (2021), Seyam et al. (2023), and Basheer et al. (2022), who applied machine learning–based classification using multi-temporal Landsat and Sentinel imagery to map land use/land cover (LULC) changes and consistently achieved high classification accuracy, generally exceeding 85%. Other studies, such as Abebe et al. (2022), Baig et al. (2022), and Girma et al. (2022) extended LULC analysis toward spatio-temporal modeling by integrating cellular automata, Markov chains, and artificial neural networks to simulate future land transformation scenarios, revealing persistent trends of forest loss, agricultural expansion, and settlement growth. In addition, Ahmad et al. (2021) and Zhu et al. (2022) linked observed LULC dynamics to environmental impacts, including groundwater quality degradation and ecosystem carbon storage reduction. Overall, these studies demonstrate the effectiveness of remote sensing and machine learning approaches for characterizing LULC dynamics, while indicating that most existing works primarily focus on spatial change patterns rather than explicitly addressing temporal consistency in multi-temporal classification results.

Compared to previous research, in this study, classification of built-up and non-built-up was carried out using Support Vector Machine (SVM), Logistic Regression (LR), and Decision Tree (DT) classification methods. This study uses two SAR sentinel images with a resolution of 10×10 so that the object used will be clearer, using two objects, namely built-up and non-built-up areas. From the results of the three methods, a comparison was made by examining the average accuracy and the smallest average variance difference in the 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 performed to determine the density proportion between non-built-up and non-built-up areas based on predetermined radii 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 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 Earth's surface. The 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 obtain the latest Earth surface scanning data (Latuamury et al., 2025). For Sentinel-2, 13 bands with their respective characteristics were installed. Four bands with spatial resolution of 10m ensured compatibility with SPOT 4/5 and meet the usability requirements for land-coverage classification. Six bands had a spatial resolution of 20m

which is a requirement for other level 2 processing parameters. The 60 m resolution band was 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 was considered sufficient to capture the spatial variability of atmospheric geophysical parameters (Huang et al., 2018; Su et al., 2021). This study specifically focuses on an area located in Umbulmartani Village, Sleman Regency, Yogyakarta, Indonesia, including its directly adjacent administrative units. Area clipping was applied to extract satellite pixels covering the Universitas Islam Indonesia (UII) campus, which is situated within Umbulmartani Village, and its surrounding regions. An illustration of the study area is presented in Figure 1.

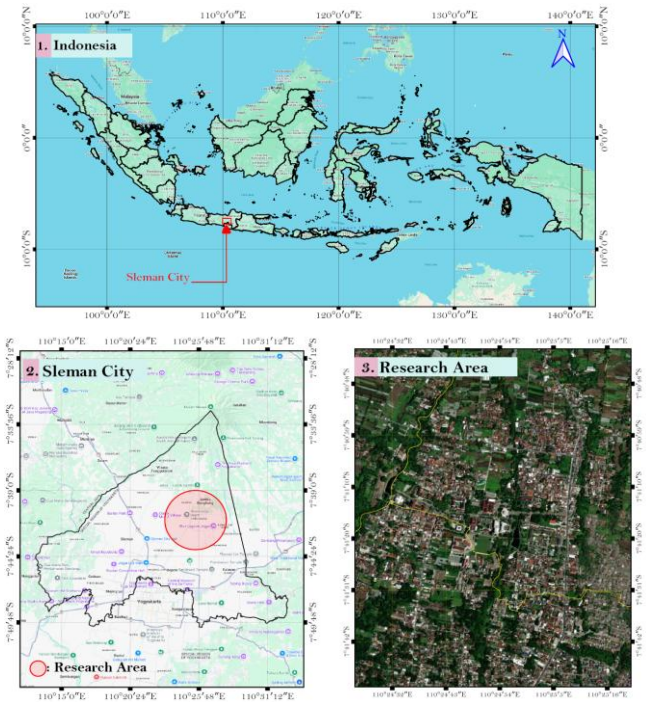

Figure 1. Research Area

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 can most effectively divide two classes. The margin value should be maximized to produce a suitable hyperplane. The margin is the separation between the support vector and hyperplane. A linear classifier is the main principle of support vector machines (Awad & Khanna, 2015; Valkenborg et al., 2023). The first step of the SVM algorithm is to define the equation of a separating hyperplane which is shown in Equation 1.

$$w.x_i + b = 0 \quad \textit{(Equation 1)}$$

where w: weight vector, x_i: i-th data, b: bias value. If the b is used as an additional weight for w_0, the following formula is used (Equation 2):

$$w_0 + w.x_i = 0 \quad \textit{(Equation 2)}$$

Equations 3 and 4 were used for every point located above or below the hyperplane used the Equation 3 and 4 (Listiana et al., 2023).

$$w_0 + w.x_i > 0 \quad \textit{(Equation 3)}$$
$$w_0 + w.x_i < 0 \quad \textit{(Equation 4)}$$

The weights are adjusted such that the hyperplane is separated into two classes. y_i is the i-th data class; therefore, Equations 5 and 6 are used.

$$H_1: w_0 + w.x_i \geq 1 \; for \; = +1 \ldots\ldots\ldots\ldots\ldots\ldots \text{(Equation 5)}$$
$$H_2: w_0 + w.x_i \geq -1 \; for \; y_i = -1 \ldots\ldots\ldots\ldots\ldots\ldots \text{(Equation 6)}$$

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 involves finding the minimum point of Equation 7, considering the constraints of Equation 8,

$$\min_w \tau(w) = \frac{1}{2}\|w\|^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(Equation 7)}$$
$$y_i(x_i.w + b) - 1 \geq 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(Equation 8)}$$

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

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

a_i is a Lagrange multiplier with a value of zero or a positive value $a_i \geq 0$. The optimal value can be calculated by minimizing L with respect to w and b. and maximizing 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.

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

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

Logistic Regression (LR) is a data analysis method used to determine 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 two categories which are usually denoted by y = 1, representing a success and y = 0, representing failure (Hagenaars et al., 2024). The probability function for every observation is given by Equation 11.

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

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

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

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

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

where π(x) is the probability of a successful event with $0 \leq \pi(x) \leq 1$, under the condition that 1 is declared a successful event and 0 indicates a failure. $\pi(x)$ is a nonlinear function; therefore, it needs to be transformed into a logit form. Parameter estimation of the logistic regression model can be described using the logit transformation of $\pi(x)$ as expressed in Equation 14. The exponential form of the logit model is presented in Equation 15.

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(Equation 14)}$$
$$\frac{\pi(x)}{1-\pi(x)} = e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)} \ldots\ldots\ldots\ldots \text{(Equation 15)}$$
$$g(x) = logit\{\pi(x)] = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \ldots \text{(Equation 16)}$$

As shown in Equation 16, g(x) is the relationship function of the LR model, referred to as the logit relationship function. The expected probability of $Y = 1$ (success) based on a given value of X is given in Equation 17.

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

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

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

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 (Fitzmaurice, 2016).

A Decision Tree (DT) classifies data items into a limited number of predetermined classes (Costa & Pedreira, 2023; Iranmanesh et al., 2022). The DT C5.0 algorithm is an improvement on Decision Tree C4.5, using a reduced tree to make the actions more succinct. The new decision tree techniques, ID3 and C4.5, are superior in terms of memory savings and pruning (Shamrat et al., 2022). 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 was performed using the gain ratio. The gain ratio measure is used to select the test attributes at each node in the tree. The attribute with the highest gain ratio is selected as the parent for the next node. The steps for creating a tree in the C5.0 algorithm are to create a tree using the C4.5 algorithm. Similarities include 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 (Abdullah et al., 2021) Equation 19 is the used to calculate the entropy value.

$$\text{Entropy S} = i{=}1n{-}pi{*}log2pi \ldots\ldots\ldots\ldots\ldots \text{(Equation 19)}$$

S: Set of cases, n: Number of partitions S, and pi: Si to S ratio. The gain value is determined as follows in Equation 20:

$$Gain\text{S},A{=}\text{Enthropy}\ S{-}i{=}1nSiS{*}\text{Entropy}Si \ldots \text{(Equation 20)}$$

A: Attribute used, $|Si|$: Number of cases in the i-th partition, $|S|$: Number of cases in S. The following is the formula used to calculate the gain ratio in Equation 21 (Priyatna et al., 2018),

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

Gain (S, A): gain value from variable, $\sum_{i=1}^{n} Entropy(Si)$: number of entropy values for 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. From the results of the training and testing, at confusion matrix was used. The Confusion Matrix (CM) provides comparative information between the classification results from the model output and 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, 2019) (see Table 1).

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 was measured using four criteria: values of accuracy, recall/sensitivity, specificity, and Area Under the Curve (AUC) from the construction of the classification model. Accuracy is the accuracy of predictions for which the model is correct. The following is the calculation for determining the accuracy value. The higher the accuracy value, the better is the classification (Tharwat, 2021; Trivedi et al., 2021; Vujovic, 2021). A recall is a ratio that is predicted to be correct or relevant when compared

to all correct data. The higher the sensitivity value, the less likely it is that the results of the positive class classification would be wrong (Vujovic, 2021). Specificity is the correctness of predicting negative data compared to overall negative data. The higher the specificity value, the better the classification performance for prediction, because it has low false positives (de Diego et al., 2022).

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 incorrectly reported as positive cases was divided by the total number of negative cases. The AUC assesses discriminatory performance by calculating the likelihood of an output from a randomly chosen sample of the positive population. The AUC values always ranged from 0 to 1. The greater the AUC value, the stronger is the classification (Egwom et al., 2022) calculation of each evaluation size, as presented in Equations 22-26,

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

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

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

$$\text{FPR} = 1 - \text{Specificity} \ldots\ldots\ldots\ldots\ldots\ldots \textit{(Equation 25)}$$

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

2.4. Research Methodology

The data used in this research were Sentinel-2 satellite image data from 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 it can be used for research. The population used was satellite imagery in an area directly adjacent to UII, and the sample used is the area directly adjacent to UII in 2021. Five variables were used in this study: variable x as Universal Transverse Mercator (UTM) easting/ longitude, variable y as UTM northing/latitude, independent variable (X), namely bands 8 and 11, and dependent variable (Y), namely class with 0 as non-built-up and 1 as a built-up. 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 (Zhou et al., 2021). These two bands were used because both are constituents of the Normalized Difference Built-Up Index (NDBI), and NDBI extraction has been carried out in previous studies (Guo et al., 2015; Kaur & Pandey, 2022; Kulkarni & Vijaya, 2021; Zheng et al., 2021).

The research flow chart is shown in Figure 2. The brown color is the first stage, namely preprocessing. Geoprocessing was carried out during the preprocessing stage. Geoprocessing is a process in a Geographic Information System (GIS) that is used to analyze spatial data, which eventually produces novel data and information. Vector geoprocessing is a geoprocessing technique implemented on vector-data structures. Examples of vector geoprocessing techniques include clipping, buffering, splitting, merging, and overlaying of vectors (Ahmed et al., 2018). The following steps were performed in the first stage. (1) Data retrieval through Copernicus Web. (2) Image cropping using SNAP. The images were cropped according to the area used. (3) Improved image resolution in band 11 and extracted band 8 using the SNAP application. The resolution increase in band 11 is achieved because band 11 has a resolution of 20 m while band 8 has a resolution of 10 m, to make the two bands have the same resolution, an increase will be made in band 11.

In the green box, data are generated that border the UII. To obtain complete data, a merger between bands 8 and 11 was carried out by clipping or cutting using a polygon area directly adjacent to the village of Umbulmartani. Next, the following steps were performed carried out in the blue-box stage. (1) Digitize built-up and non-built-up areas by creating polygons in the Quantum Geographic Information System (QGIS). Digitization is the process of converting an analog map into a digital map by using a digitization table. The method converts the existing spatial features on the map into a set of x, y coordinates. High-quality analog map

sources are required to produce accurate data. For the digitization process, the operator requires high precision and concentration are required from the operator (Cunningham, 2006), (2) Sample the training and testing data by conducting random points in areas that have been digitized. (3) Classification analysis using support vector machines, logistic regression, and decision tree methods with 40 iterations. Iteration was used to determine the pattern and stability of the predictions obtained. (4) Descriptive analysis was performed using the iteration results of each classification. (5) The methods were compared using the average indicator of each method to obtain the best method.

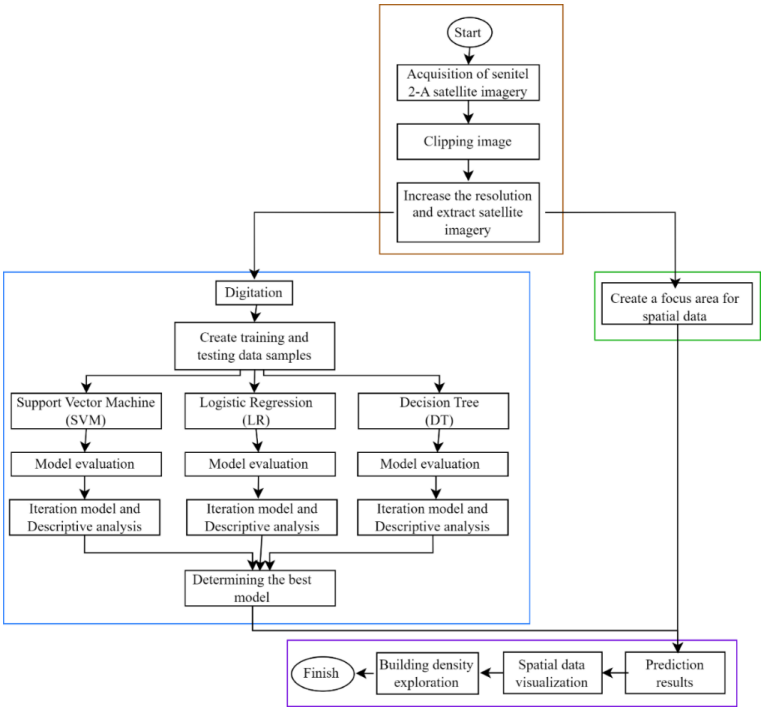

Figure 2. Research Flowchart

In the last stage (purple box), the following steps were performed. (1) Predict the best methods and datasets to obtain predictive images using raster visualization. (2) A buffering analysis was performed to calculate the proportion of built-up and non-built-up areas. Buffering refers to the creation of zones or areas using 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). The buffers are polygonal in shape because they represent the area around an object. Buffering also refers to the formation of sonas or corridors in raster data models (Li, 2017). R software is used for classification most use R software because of its several advantages, such as convenience, portability, multiplatform, and programmability (Giorgi et al., 2022; Pavlenko et al., 2020). The stages of spatial data visualization and preprocessing are used by QGIS, because it 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; Löwe et al., 2022; Rosas-Chavoya et al., 2022).

3. Result and Discussion

3.1. Data Preparation

Area clipping was performed to obtain the study area, which is directly adjacent to 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 have 524,521 points with four 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 training and testing data were prepared by obtaining random points from the digitized results. The total number of data was 2000 which consisted of 1000 built-up and 1000 non-built-up areas. Figure 3 shows the digitization results for the built-up and non-built-up areas. The green polygons represent built-up, and the brown polygons represent non-built-up areas. Digitization was carried out to obtain training and testing data that were labeled 0 for non-built-up and 1 for built-up. We split the data into 80 percent training data with 1600 rows and 20 percent testing data with 400 rows.

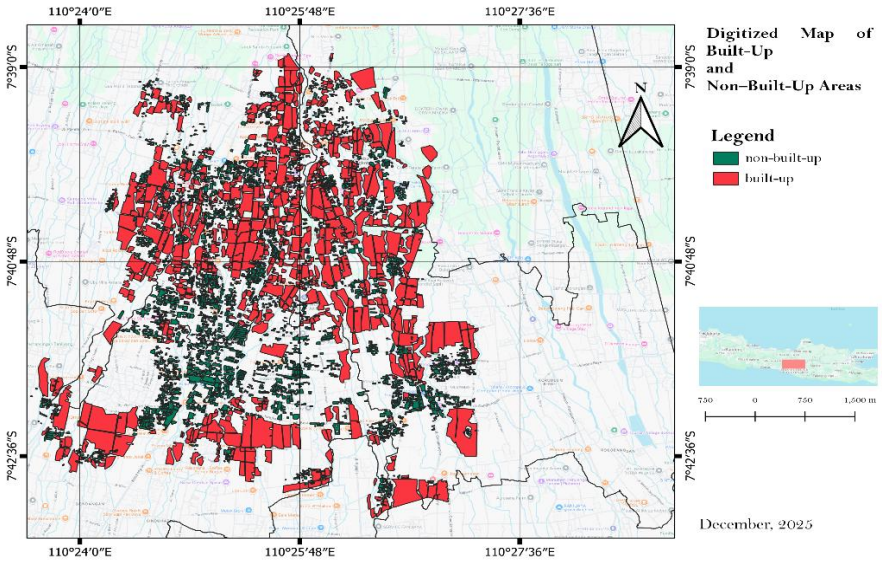

Figure 3. Digitization

3.2. Classification

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

Table 4. Evaluation Results of Support Vector Machine Model

Iteration	Training				Testing			
	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	89.8	90.2	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 is carried out in the logistic regression method, based on Equation 14, and the resulting form of the logistic regression equation is presented in Equation 27.

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = -1.486 - 0.004X_1 + 55.201\,X_2 \quad \ldots\ldots \textit{(Equation 27)}$$

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

The calculated odds ratio value is presented in Equation 28.

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

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 are included in Y=0 (non-built-up). Table 5 shows the values of accuracy, sensitivity, specificity, and AUC for each iteration using the training and testing data. Table 5 presents the results of the 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	91.5	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

Table 6 shows the values of accuracy, sensitivity, specificity, and AUC for the DT method classification.

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	89.6	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 from the training and testing data, as well as three other indicators based on the number of iterations accomplished (sensitivity, specificity, and AUC), as 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 a comparison of the mean values from all classifications we found that the decision tree model has the highest accuracy. However, the SVM model had the smallest difference in mean values between

classifications on the training and test data, which was 0.06. In addition, it has a relatively small difference in variance compared with the other methods. Therefore, we declared SVM to be 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 (SVM), into intact data. 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 points with longitude, latitude, and class variables as the predicted outcome variables.

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

The prediction results obtained were used as the raster data. The results obtained from the predictions portray non-built-up in white and built-up in red, as shown in Figure 4.

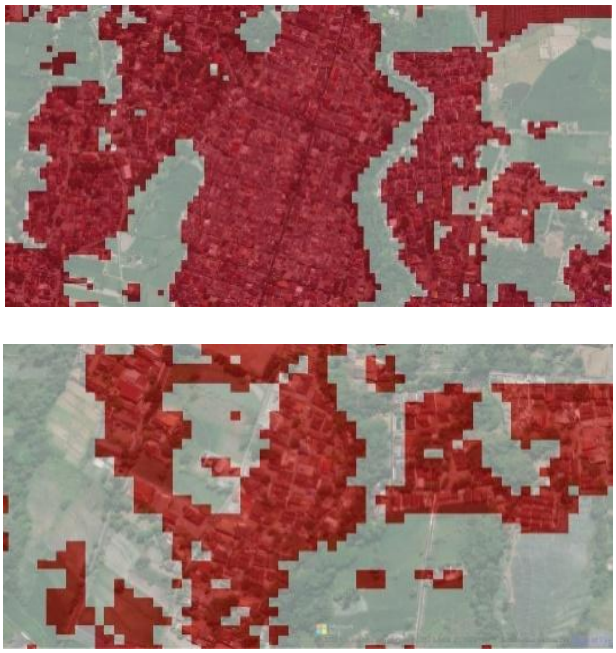

Figure 4. Prediction Results.

Figure 4 shows the predicted results when zoomed-in. In conclusion, the predictions are in accordance with the original base map. After the classification and prediction results were obtained, one of its uses was to calculate the proportions.

3.5. Proportional Calculation

A buffering analysis was performed to identify the relationship between a point and the surrounding area. In this study we drew a buffer with a center point at the UII and zones with radii of 1, 2, 3, 4, and 5 km are made.

Figure 5 show the results for the buffer. Buffers are obtained using a distance of 1-5 km where Figure 5 show an example of the points that are in the buffer. The yellow dots are the built-up points, whereas the blue dots are the non-built-up points. A buffer analysis of the complete data is presented in Table 9.

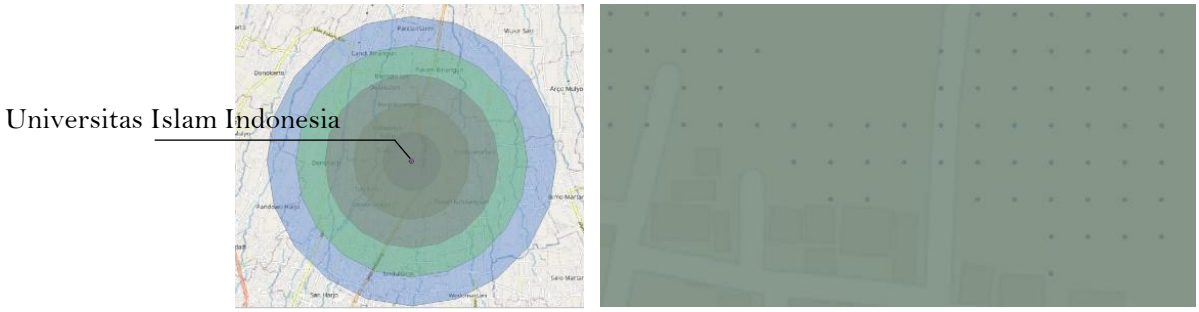

Figure 5. Point Spatial Buffer from UII

Based on Table 9, it was found that at a radius of 1 km the percentage of built-up was 47.695% while the remaining 52.305% were non-built-up. At a radius of 2 km, the percentage of built-up was 39,737% while the remaining 60,263% were non-built-up. At a radius of 3 km, the percentage of built-up was 33,271% whereas 66,729% of the points were non-built-up. At a radius of 4 km, the percentage of built-up was 31,597% whereas 68,403% of the points were non-built-up. At a radius of 5 km, the percentage of built-up was 30.597% whereas 69.403% of the points were non-built-up. We arrived at the conclusion that the closer to the central point, the greater the percentage of built-up, while for non-built-up, the greater the distance, the greater the percentage. This suggests that the UII campus area serves as a new residential center. Many built-up areas are erected to support services for the academic community in addition to entrepreneurs who want to create a business that requires premises that often target areas close to UII to open their businesses. Areas with dense populations increase community activities that cause traffic congestion, high water demand, large garbage production, high production of water waste, plastic waste, and so one. Therefore, it is urgent that the relevant agencies carry out rearrangement and maintenance to reduce the impact of the establishment of many non-built-up areas.

Table 9. Buffering Analysis

Distance (km)	Area (km²)	Built-up Points	Non-built-up Points	Percentage of Built-up (%)	Percentage of Non-Built-up (%)
1	3,09	14,73	16,16	47.69	52.30
2	12,37	46,03	69,80	39.73	60.26
3	27,83	79,54	159,53	33.27	66.72
4	49,48	112,77	244,12	31.59	68.40
5	77,30	136,24	309,03	30.59	69.40

These findings are consistent with recent studies demonstrating that university campuses often act as growth poles that stimulate land-use transformation in their surrounding areas. Çalışkan (2023), found that the establishment of new universities in Turkey accelerated the expansion of built-up areas and reduced nearby green spaces. Similarly, Widaningrum (2022) reported that educational facilities tend to trigger the development of supporting infrastructure and commercial activities, thereby increasing the extent of built-up zones. In addition, Dharmawan et al. (2021) showed that the physical expansion of university campuses in Indonesia significantly reshaped the local landscape by increasing the proportion of impervious surfaces relative to green areas. In the context of the UII campus, these results confirm that the campus functions not only as an academic hub but also as a driver of urban formation. The spatial pattern observed, where higher built-up proportions occur within 1–2 km and gradually decrease with greater distance, illustrates a typical process of urban diffusion driven by accessibility, land value, and service demand. Therefore, local governments and urban planners should anticipate

continuing urban pressure by implementing appropriate zoning policies, encouraging vertical development, and preserving green buffers to maintain sustainable growth around the campus (Ślawska, 2024).

4. Conclusion

Based on the iteration results, the average values for accuracy, sensitivity, specificity, and AUC in the training data, classified using the SVM method, were 91.3073%, 90.7990%, 91.83%, 91.3125% respectively, whereas in the testing data, they were 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 were, 91.2630%, 90.7988%, 91.7563%, 91.2750% respectively whereas in the testing data they were 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, were 91.9250%, 92.1365%, 91.7845%, 91.9675%, respectively, while those of the testing data were 90.4375%, 91.0190 respectively. %, 89.9415%, 90.4850%. The best method used in this study was the SVM method which was determined based on the average accuracy and the smallest average variance difference between the training and testing data, and three other indicators based on 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 of built-up was 47.695% while for non-built-up it was 52.305%. At a radius of 2 km, a percentage of 39,737% was obtained for built-up while for non-built-up it was 60,263%. At a radius of 3 km, the percentage of built-up was 33,271% while that of non-built-up was 66,729%. At a radius of 4 km, the percentage of built-up was 31,597%, whereas the remaining 68,403% of the points were non-built-up. At a radius of 5 km, the percentage of built-up was 30.597% while the remaining 69.403% were non-built-up. We conclude that for both datasets, the closer to the center point, the greater the percentage of built-up. In contrast, the farther away from the center point, the greater the percentage of non-built-up areas.

5. Acknowledgments

The authors would like to express their sincere gratitude to the Directorate of Research and Community Service (Direktorat Penelitian dan Pengabdian Masyarakat/DPPM), Universitas Islam Indonesia, for providing financial support through a research grant under Contract No. 03/Dir/DPPM/70/Pen.Pemula/III/2020. The authors also acknowledge the research facilities and institutional support made available by Universitas Islam Indonesia. Ethical approval was not required for this study, as it did not involve human participants, human data or tissue, or the use of animals. The study relied solely on publicly available satellite imagery and secondary geospatial datasets. The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

6. References

Abdullah, A. Z., Winarno, B., & Saputro, D. R. S. (2021). The Decision Tree Classification with C4.5 and C5.0 Algorithm based on R to Detect Case Fatality Rate of Dengue Hemorrhagic Fever in Indonesia. *Journal of Physics: Conference Series*, *1776*(1), 012040. [Crossref]

Abebe, G., Getachew, D., & Ewunetu, A. (2022). Analysing Land Use/Land Cover Changes and Its Dynamics Using Remote Sensing and GIS in Gubalafito District, Northeastern Ethiopia. *SN Applied Sciences*, *4*(1), 30. [Crossref]

Ahmad, W., Iqbal, J., Nasir, M. J., Ahmad, B., Khan, M. T., Khan, S. N., & Adnan, S. (2021). Impact of Land use/Land Cover Changes on Water Quality and Human Health in District Peshawar Pakistan. *Scientific Reports*, *11*(1), 16526. [Crossref]

Ahmed, Z., Krupnik, T. J., & Kamal, M. (2018). Introduction to Basic GIS and Spatial Analysis Using QGIS: Applications in Bangladesh. *Dhaka, CIMMYT--Bangladesh*, 80–94.

Athanasiou, L. S., Fotiadis, D. I., & Michalis, L. K. (2017). Propagation of Segmentation and Imaging System Errors. In *Atherosclerotic Plaque Characterization Methods Based on Coronary Imaging* (pp. 151–166). Elsevier. [Crossref]

Awad, M., & Khanna, R. (2015). Support Vector Machines for Classification. In *Efficient Learning Machines* (pp. 39–66). Apress. [Crossref]

Baig, M. F., Mustafa, M. R. U., Baig, I., Takaijudin, H. B., & Zeshan, M. T. (2022). Assessment of Land Use Land Cover Changes and Future Predictions Using CA-ANN Simulation for Selangor, Malaysia. *Water*, *14*(3), 402. [Crossref]

Basheer, S., Wang, X., Farooque, A. A., Nawaz, R. A., Liu, K., Adekanmbi, T., & Liu, S. (2022). Comparison of Land Use Land Cover Classifiers Using Different Satellite Imagery and Machine Learning Techniques. *Remote Sensing*, *14*(19), 4978. [Crossref]

Bettinger, P., Boston, K., Siry, J. P., & Grebner, D. L. (2017). Geographic Information and Land Classification in Support of Forest Planning. In *Forest Management and Planning* (pp. 65–85). Elsevier. [Crossref]

Bruy, A., & Svidzinska, D. (2015). *QGIS by Example*. Packt Publishing Ltd.

Çalışkan, E. B. (2023). Tracking the University Campus Development by Using Remote Sensing and Satellite Imagery. *Gazi University Journal of Science Part B: Art Humanities Design and Planning*, *11*(2), 237–252.

Costa, V. G., & Pedreira, C. E. (2023). Recent Advances in Decision Trees: An Updated Survey. *Artificial Intelligence Review*, *56*(5), 4765–4800. [Crossref]

Cunningham, M. A. (2006). Accuracy Assessment of Digitized and Classified Land Cover Data for Wildlife Habitat. *Landscape and Urban Planning*, *78*(3), 217–228.[Crossref]

de Diego, I. M., Redondo, A. R., Fernández, R. R., Navarro, J., & Moguerza, J. M. (2022). General Performance Score for Classification Problems. *Applied Intelligence*, *52*(10), 12049–12063. [Crossref]

Dharmawan, I. A., Rahadianto, M. A. E., Henry, E., Endyana, C., & Aufaristama, M. (2021). Application of High-Resolution Remote-Sensing Data for Land Use Land Cover Mapping of University Campus. *The Scientific World Journal*, *2021*, 1–17. [Crossref]

Egwom, O. J., Hassan, M., Tanimu, J. J., Hamada, M., & Ogar, O. M. (2022). An LDA–SVM Machine Learning Model for Breast Cancer Classification. *BioMedInformatics*, *2*(3), 345–358. [Crossref]

Fitzmaurice, G. M. (2016). Regression. *Diagnostic Histopathology*, *22*(7), 271–278. [Crossref]

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. [Crossref]

Giorgi, F. M., Ceraolo, C., & Mercatelli, D. (2022). The R Language: An Engine for Bioinformatics and Data Science. *Life*, *12*(5), 648. [Crossref]

Girma, R., Fürst, C., & Moges, A. (2022). Land Use Land Cover Change Modeling by Integrating Artificial Neural Network with Cellular Automata-Markov Chain Model in Gidabo River Basin, Main Ethiopian Rift. *Environmental Challenges*, *6*, 100419. [Crossref]

Government of the Republic of Indonesia (2002) Law No. 28 of 2002 on Buildings. Indonesia

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. [Crossref]

Hagenaars, J. A. P., Kühnel, S., & Andress, H.-J. (2024). *Interpreting and Comparing Effects in Logistic,Probit, and Logit Regression* (Issue 194). Sage Publications.

Huang, H., Sun, G., Ren, J., Rang, J., Zhang, A., & Hao, Y. (2018). Spectral-Spatial Topographic Shadow Detection from Sentinel-2A MSI Imagery Via Convolutional Neural Networks. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, 661–664. [Crossref]

Iranmanesh, M., Seyedabrishami, S., & Moridpour, S. (2022). Identifying High Crash Risk Segments in Rural Roads Using Ensemble Decision Tree-based Models. *Scientific Reports*, *12*(1), 20024. [Crossref]

Janipella, R., Gupta, V., & Moharir, R. V. (2019). Application of Geographic Information System in Energy Utilization. In *Current Developments in Biotechnology and Bioengineering* (pp. 143–161). Elsevier. [Crossref]

Kaur, R., & Pandey, P. (2022). A Review on Spectral Indices for Built-Up Area Extraction Using Remote Sensing Technology. *Arabian Journal of Geosciences*, *15*(5), 391. [Crossref]

Kulkarni, K., & Vijaya, P. (2021). NDBI Based Prediction of Land Use Land Cover Change. *Journal of the Indian Society of Remote Sensing*, *49*(10), 2523–2537. [Crossref]

Latuamury, B., Imlabla, W. N., Sahusilawane, F., & Marasabessy, H. (2025). Spatial Patterns of NDVI Vegetation Greenness Index and NDWI Water Content Index Using Sentinel 2A Imagery. *IOP Conference Series: Earth and Environmental Science*, *1527*(1), 012002. [Crossref]

Li, X. (2017). Buffers. *Geographic Information Science & Technology Body of Knowledge*, *2017*(Q4). [Crossref]

Listiana, E., Muzayanah, R., Muslim, M. A., & Sugiharti, E. (2023). Optimization of Support Vector Machine Using Information Gain and AdaBoost to Improve Accuracy of Chronic Kidney Disease Diagnosis. *Journal of Soft Computing Exploration*, *4*(3), 152–158. [Crossref]

Löwe, P., Anguix Alfaro, Á., Antonello, A., Baumann, P., Carrera, M., Durante, K., Hugentobler, M., Lime, S., Mitasova, H., Müller, D., Neteler, M., Reed, J., Strobl, C., & Wessel, P. (2022). *Open Source – GIS* (pp. 807–843). [Crossref]

Nugroho, A. S., Witarto, A. B., & Handoko, D. (2003). Support Vector Machine. *Proceeding Indones. Sci. Meeiting Cent. Japan*.

Pavlenko, L., Pavlenko, M., Khomenko, V., & Mezhuyev, V. (2020). Application of R Programming Language in Learning Statistics. *Proceedings of the 1st Symposium on Advances in Educational Technology*, 62–72. [Crossref]

Pebesma, E., & Bivand, R. (2023). *Spatial Data Science: With Applications in R*. Chapman and Hall/CRC.

Priyatna, R. D., Tulus, & Ramli, M. (2018). K-Means Algorithm and Modification Using Gain Ratio. *IOP Conference Series: Materials Science and Engineering*, *420*, 012133. [Crossref]

Rosas-Chavoya, M., Gallardo-Salazar, J. L., López-Serrano, P. M., Alcántara-Concepción, P. C., & León-Miranda, A. K. (2022). QGIS a Constantly Growing Free and Open-Source Geospatial Software Contributing to Scientific Development. *Cuadernos de Investigación Geográfica*, *48*(1), 197–213. [Crossref]

Seyam, M. M. H., Haque, M. R., & Rahman, M. M. (2023). Identifying the Land Use Land Cover (LULC) Changes Using Remote Sensing and GIS Approach: A Case Study at Bhaluka in Mymensingh, Bangladesh. *Case Studies in Chemical and Environmental Engineering*, *7*, 100293. [Crossref]

Shamrat, F. M. J. M., Ranjan, R., Hasib, K. M., Yadav, A., & Siddique, A. H. (2022). *Performance Evaluation Among ID3, C4.5, and CART Decision Tree Algorithm* (pp. 127–142). [Crossref]

Ślawska, J. (2024). Shaping the Rural Landscape: Institutions of Land Use Change in Non-Urbanized Areas in Poland. *Sustainability*, *16*(24), 10902. [Crossref]

Su, H., Peng, Y., Xu, C., Feng, A., & Liu, T. (2021). Using Improved DeepLabv3+ Network Integrated with Normalized Difference Water Index to Extract Water Bodies in Sentinel-2A Urban Remote Sensing Images. *Journal of Applied Remote Sensing*, *15*(01). [Crossref]

Syahrani, I. M. (2019). Comparation Analysis of Ensemble Technique With Boosting(Xgboost) and Bagging (Randomforest) For Classify Splice Junction DNA Sequence Category. *Jurnal Penelitian Pos Dan Informatika*, *9*(1), 27–36. [Crossref]

Tharwat, A. (2021). Classification Assessment Methods. *Applied Computing and Informatics*, *17*(1), 168–192. [Crossref]

Trivedi, N. K., Gautam, V., Anand, A., Aljahdali, H. M., Villar, S. G., Anand, D., Goyal, N., & Kadry, S. (2021). Early Detection and Classification of Tomato Leaf Disease Using High-Performance Deep Neural Network. *Sensors*, *21*(23), 7987. [Crossref]

Valkenborg, D., Rousseau, A.-J., Geubbelmans, M., & Burzykowski, T. (2023). Support Vector Machines. *American Journal of Orthodontics and Dentofacial Orthopedics*, *164*(5), 754–757. [Crossref]

Vujovic, Ž. Đ. (2021). Classification Model Evaluation Metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6). [Crossref]

Widaningrum, D. L. (2022). The Effect of Supporting Facilities Growth Around the Urban Campus on Land-Use Change. *Proceedings of the International Conference on Engineering and Information Technology for Sustainable Industry*, 1–6. [Crossref]

Yao, X., & Li, G. (2018). Big Spatial Vector Data Management: A Review. *Big Earth Data*, *2*(1), 108–129. [Crossref]

Zhai, H., Lv, C., Liu, W., Yang, C., Fan, D., Wang, Z., & Guan, Q. (2021). Understanding Spatio-Temporal Patterns of Land Use/Land Cover Change under Urbanization in Wuhan, China, 2000–2019. *Remote Sensing*, *13*(16), 3331. [Crossref]

Zheng, Y., Tang, L., & Wang, H. (2021). An Improved Approach for Monitoring Urban Built-Up Areas by Combining NPP-VIIRS Nighttime Light, NDVI, NDWI, and NDBI. *Journal of Cleaner Production*, *328*, 129488. [Crossref]

Zhou, J., Qiu, Y., Chen, J., & Chen, X. (2021). A Geometric Misregistration Resistant Data Fusion Approach for Adding Red-Edge (RE) and Short-Wave Infrared (SWIR) Bands to High Spatial Resolution Imagery. *Science of Remote Sensing*, *4*, 100033. [Crossref]

Zhu, L., Song, R., Sun, S., Li, Y., & Hu, K. (2022). Land Use/Land Cover Change and Its Impact on Ecosystem Carbon Storage in Coastal Areas of China from 1980 to 2050. *Ecological Indicators*, *142*, 109178. [Crossref]