Computer model for detecting tsunami wave hazard on built-up land using machine learning and sentinel 2A satellite imagery

Sri Yulianto Joko Prasetyo[1], Wiwin Sulistyo[1], Erwin Christanto[1], Bistok Hasiholan Simanjuntak[2]
[1]Department of Informatic Engineering, Information Technology Faculty, Satya Wacana Christian University, Salatiga, Indonesia
[2]Bussines and Agriculture Faculty, Satya Wacana Christian University, Salatiga, Indonesia

Article Info

Article history:

Received Maret 5, 2023
Revised Sep 20, 2023
Accepted Nov 7, 2023

Keywords:

Build-up
Computer model
Machine learning
Spatial interpolation
Tsunami
Vegetation indices

ABSTRACT

The aim of this research is to compile a tsunami wave hazard scale based on built-up land density extracted and classified by machine learning from Sentinel 2A satellite and digital elevation model (DEM) imageries. This research was carried out in 5 stages, namely: (i) pre-processing of Sentinel 2A and DEM images, (ii) Classification of VI data using the machine learning algorithms, (iii) Spatial prediction using the ordinary kriging method, (iv) Field testing using the confusion matrix method, (v) Preparation of decision matrix for tsunami wave hazard. The results of the study show that the most accurate classification algorithm for classifying built-up indices data is the k-nearest neighbor (k-NN) algorithm. The results of the statistical accuracy test show that the most accurate is normalized difference built-up index (NDBI) with a mean of square error (MSE) value of 0.073 and a mean of absolute error (MAE) of 0.003. DEM analysis shows that the research area is at an altitude of 0–15 meters above sea level so it is in the high vulnerability to medium vulnerability category. Field testing showed user accuracy of 91.11%, manufacturer accuracy of 92.16%, and overall average accuracy of 91%.

Corresponding Author:

Sri Yulianto Joko Prasetyo
Department of Informatic Engineering, Information Technology Faculty
Satya Wacana Christian University
Salatiga, Indonesia
Email: sri.yulianto@uksw.edu

1. INTRODUCTION

Machine learning (ML) is a type of artificial intelligence that uses certain algorithms through the process of analyzing large amounts of data which their nature is multidimensional and uses patterns to produce new values as predictive data based on the provided historical data features [1]–[3]. ML is an algorithm for the process of predicting and detecting a phenomenon that is described in remote sensing imagery and occurs in locations or positions that are not accessible to human vision, namely the support vector machine (SVM), random forest (RF), k-nearest neighbour (k-NN), multivariate adaptive regression splines (MARS) and artificial neural networks (ANN) [4]–[6]. ML is used as a method for computing, classifying and predicting data in the form of pixels or digital number (DN) derived from Landsat 8 OLI and Sentinel 2A satellite images [7]–[10]. The research on the classification of buildings destroyed as the impact of tsunami wave in Japan in 2011 and the impact of earthquake and tsunami wave in 2016 was done using the SVM algorithm [11]. Prediction of inundation on coastal tsunami wave is made using the ANN algorithm [12]. Tsunami wave vulnerability modelling on the coast and residential areas is created using data from normalized difference vegetation index (NDVI), modified soil adjusted vegetation index (MSAVI), normalized difference water index (NDWI), modified normalized difference water index (MNDWI), and normalized difference built-up index (NDBI) extracted from Sentinel 2A satellite images. The Sentinel-2A MSI images has 13 spectral bands in the visible, NIR, and SWIR wavelength region with spatial resolutions of 10–60 m as shown in Table 1 [13]. The SVM aims to achieve optimal separation between hyper planes and/or hyper planes located in height and/or hyper planes in dimensional space, in order to find the optimal boundaries between classes. The SVM formula is shown as (1) [14],

$$f(x) = \sum_{i-1}^{n}(\alpha_i - \alpha_j)K(N_i - N_j) + c \tag{1}$$

The (1) shows that $K\left(N_i - N_j\right)$ is a kernel function which is used as a higher dimension transformation from non-linier function to linier function. The RF algorithm is called the ensemble learning method because the first decision making on a tree is made by a random subspace method, and the second data classification is made by a stochastic discriminant method [15]–[17]. RF is an ensemble learning-based algorithm which has an advantage in its resistance to noise in large data sets. Each input data will form a class in the form of a tree classification, and RF is able to form a tree classification according to the size of the input data in the form of numeric data, pixel data and spectrum data of satellite images. The RF uses Gini index to select a tree classification to produce a decision. $G_{gini}(D)$ Index represents the uncertainty of the sum of VI values as the sample of this study. Gini Index is defined as (2) [18],

$$\mathrm{G_{gini}(D)} = 1 - \sum_{\mathrm{n-1}}^{\mathrm{N}}\left(\frac{|\mathrm{A_n}|}{\mathrm{D}}\right)^2 \tag{2}$$

Where A_n is the sum of data obtained from the results of the observation in n class. An ANN algorithm is composed of 3 layers, namely input layer, hidden layer and output layer. The pixel, numeric or spectrum data extracted from the satellite image is transferred from the input layer to the output layer via neurons. The value of the connection weight between nodes is determined randomly, then the differences between the actual weight value and the predicted value are computed so that the weight value of the computational results gets closer to the actual condition. The parameters in the ANN are the number of layers, the number of neurons, learning algorithms and the activation function. The formula for data processing on ANN neurons is as (3),

$$y(t) = F(\sum_{i=0}^{m} b_i(t) . x_i(t) + c) \tag{3}$$

where $x_i(t)$ is the input value of pixel, numeric or spectrum data of satellite imagery, with the value of x is in discrete form with a value from of i of 0 to m. The value of $b_i(t)$ is the weight value in discrete form with a value of i from of 0 to m. The value of c is the bias, the value of F is the transfer function from one neuron to the next neuron, and the value of $y(t)$ is the final value in discrete form [19]. The concept of the k-NN algorithm works by identifying pixel, numeric or spectrum data that are not recognized to be included as members of the pixel, numeric or spectrum data class of the nearest recognized class [20]. Suppose a training data set is denoted as $D = \{(x_1 - y_1)(x_2 - y_2)(x_3 - y_3) \ldots (x_n - y_n)\}$, the number of training data variances or training parameters is n , vector data is denoted as $x_i \in \Re^d$ and $y_i \in f = (a_1, a_2, a_3 \ldots a_n)$ with the value of i = 1,2,3…N. The distance of the analyzed sample is calculated from the training data as the value of k-NN which is denoted by $S_k\ (x)$. The neighborhood relationship between input data and training data is $M_l = EN_l$ where the value of M_l is the set of neighbors, the value of E is the training data and N_l is the matrix coefficient. The matrix coefficient value is calculated using (4) [21], [22].

$$\overline{N_l} = argmin\|M_l - EN_l\|_F^2, s.t. \|N_l\|_{row,0} \leq t_0 \tag{4}$$

The definition of the distance between the sample value and the class value is calculated using (5).

$$d(M_l; E^x) = \|M_l - E^m\overleftarrow{N_l^{\,m}}\|_F^2 \tag{5}$$

MARS is the method used to analyze various problems and natural phenomena that always leads to a high dimensional nonlinear orientation using a linear approach in different intervals [23]. The MARS equation is as (6).

$$y = \beta_0 + \sum_{i=1}^{M}\beta_i \mathrm{B}_i(X) + \varepsilon \tag{6}$$

In which β_0 is the intercept of the model, $B_i(X)$ is a linear basis function pair where the value (X) is the vector of independent variables. The notation of β_i is the i^{th} coefficient, the M notation is the number of basic functions, and ε is the error rate [24].

This research is focused on solving the following problems: (i) Identifying and predicting distribution patterns of built-up land objects in tsunami wave hazard areas from satellite imagery data, (ii) Making an accurate classification of the tsunami wave hazard level on built-up land objects, iii) Determining of decision matrix and tsunami wave hazard scale from the best algorithm, (iv) Testing of accuracy using Confusion Matrix. In accordance with the research problems, the proposed solutions are: (i) Extracting built-up land objects from Sentinel 2A imageries using the algorithms of VI, (ii) Conducting experiments on selecting a classification method using the ML algorithms, (iii) Building a computer model architecture for an effective and efficient process of extracting, interpreting and classifying tsunami wave hazard levels, (iv) Conducting tests to identify the accuracy in the field. The novelty of this research is the existence of a computer-based model and a simulation to carry out comparisons and selection of tsunami wave hazard classifications using ML algorithms based on built-up land objects using VI. This research refines previous research that uses ML algorithms of ANN, RF, SVM, MARS, CART, k-NN, and LASSO [20].

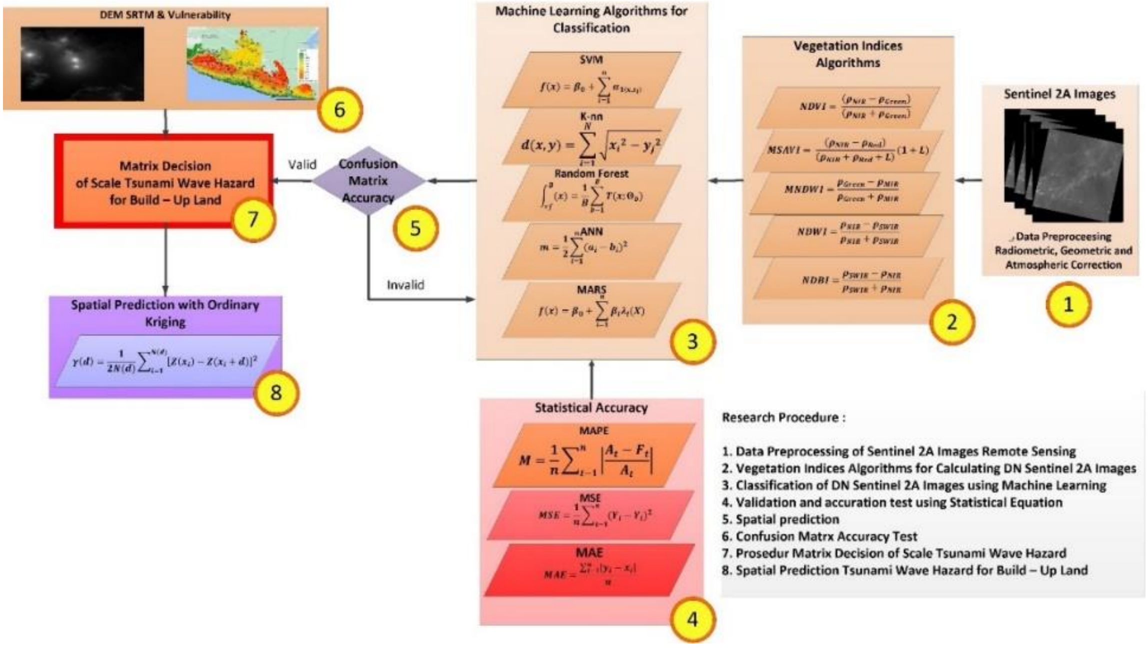
Table 1. Vegetation Indices NDVI, MSAVI, NDWI, MNDWI, and NDBI

Description	Equation		Reference
NDVI is the algorithm to study the growth and health of plants in relation to climatic and seasonal factors. NDVI is calculated using the spectrum of Red (ρred) and Near Infrared (NIR) (ρNIR) lights is as follows.	NDVI = $\frac{\rho_{NIR} - \rho_{red}}{\rho_{NIR} + \rho_{red}}$	(7)	[25], [26]
NDWI and MNDWI are the algorithms used to detect water bodies on the earth's surface, such as rivers, lakes, reservoirs and beaches macroscopically, in real-time, dynamically and efficiently compared to conventional measurements on lakes is as follows.	$NDWI = \frac{\rho_{green} - \rho_{NIR}}{\rho_{green} + \rho_{NIR}}$	(8)	[27]
MNDWI is calculated using the spectrum of Green (ρgreen) and Middle Infra-Red (MIR) (ρMIR) light. NDWI is calculated using the spectrum of Green (ρGreen) and Near Infrared (NIR) (ρNIR) lights, and the NIR spectrum in MNDWI is replaced by Shortwave Infrared (SWIR) (ρSWIR) light is as follows.	$MNDWI = \frac{\rho_{green} - \rho_{MIR}}{\rho_{green} + \rho_{MIR}}$	(9)	[27]
NDBI is the algorithm used to study the density of built-up lands, using the spectrum of Shortwave Infrared (SWIR) (ρSWIR) and Near Infrared (NIR) (ρNIR) lights is as follows.	$NDBI = \frac{\rho_{SWIR} - \rho_{NIR}}{\rho_{SWIR} + \rho_{NIR}}$	(10)	[28]
MSAVI is the improved algorithm of NDVI by reducing the reflectance factor of the soil background in order to produce a more accurate vegetation cover is as follows.	$MSAVI = \frac{\rho_{NIR} - \rho_{red}}{\rho_{NIR} + \rho_{red} + L_o}(1 + L_o)$	(11)	[29]

2. METHOD

The research location is in Gunungkidul Regency, Special Region of Yogyakarta Province, Indonesia which includes 78 villages, in an area of 8 sub-districts. Geographically, the study area is located at coordinates of 7° 46' - 8° 09' South Latitude and 110° 21' - 110° 50' East Longitude. The research area is classified into 4 parts, namely: built-up land consisting of settlements, traditional markets and road networks (indicated by yellow color), agricultural land consisting of dry fields, rice fields and gardens, forest and mixed vegetation consisting of open land, grasses and shrubs. This research uses imageries obtained from the Sentinel 2A satellite. The research was carried out in 4 stages. The first stage is pre-processing stage consisting of geometric, radiometric, atmospheric corrections and extraction of satellite images using the vegetation indices algorithms (Figure 1). Next is data extraction using NDVI, MSAVI, MNDWI, NDWI, NDBI algorithms. This stage aims to transform data from image data with pixel components into numerical form of vegetation indices data. The second stage is classifying VI data using ML algorithms namely SVM, k-NN, RF, ANN and MARS. Accuracy testing was carried out using statistical methods, namely MAE, MSP and MAPE.

The third stage is to make spatial predictions at locations that are not observation area using the ordinary kriging method. The fourth stage is field testing using the confusion matrix method to see the suitability and accuracy of the computer models generated from computations with machine learning with the real conditions at the Gunungkidul Regency, Yogyakarta, Indonesia. The fifth stage is to analyze the SRTM DEM data to see the elevation of the observation area.


DEM SRTM & Vulnerability
6
Matrix Decision of Scale Tsunami Wave Hazard for Build – Up Land
7
Valid
Confusion Matrix Accuracy
5
Invalid
Spatial Prediction with Ordinary Kriging
$\gamma(d) = \frac{1}{2N(d)}\sum_{i=1}^{N(d)}[Z(x_i) - Z(x_i + d)]^2$
8
Machine Learning Algorithms for Classification
SVM
$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_{1(x,x_j)}$
K-nn
$d(x,y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_j^2}$
Random Forest
$\int_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\Theta_b)$
ANN
$m = \frac{1}{2}\sum_{i=1}^{n}(a_i - b_i)^2$
MARS
$f(x) - \beta_0 + \sum_{i=1}^{n} \beta_i \lambda_i(X)$
3
Vegetation Indices Algorithms
$NDVI = \frac{(\rho_{NIR} - \rho_{Green})}{(\rho_{NIR} + \rho_{Green})}$
$MSAVI = \frac{(\rho_{NIR} - \rho_{Red})}{(\rho_{NIR} + \rho_{Red} + L)}(1 + L)$
$MNDWI = \frac{\rho_{Green} - \rho_{MIR}}{\rho_{Green} + \rho_{MIR}}$
$NDWI = \frac{\rho_{NIR} - \rho_{SWIR}}{\rho_{NIR} + \rho_{SWIR}}$
$NDBI = \frac{\rho_{SWIR} - \rho_{NIR}}{\rho_{SWIR} + \rho_{NIR}}$
2
Sentinel 2A Images
Data Preproceesing Radiometric, Geometric and Atmospheric Correction
1
Statistical Accuracy
MAPE
$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$
MSE
$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
MAE
$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$
4

Research Procedure :

1. Data Preprocessing of Sentinel 2A Images Remote Sensing
2. Vegetation Indices Algorithms for Calculating DN Sentinel 2A Images
3. Classification of DN Sentinel 2A Images using Machine Learning
4. Validation and accuration test using Statistical Equation
5. Spatial prediction
6. Confusion Matrx Accuracy Test
7. Prosedur Matrix Decision of Scale Tsunami Wave Hazard
8. Spatial Prediction Tsunami Wave Hazard for Build – Up Land

Figure 1. Comparison and selection of the computer model for tsunami wave hazard classification using the ML algorithms based on built-up land objects using VI information

3. RESULTS AND DISCUSSION

Based on the results of previous research, on the south coast of Central Java and Jogjakarta, especially the areas that are highly vulnerable to tsunami waves, land cover consists of: (i) residential buildings for socio-economic activities and tourism, (ii) agricultural lands, plantations and forests, (iii) aquaculture ponds, and (iv) open grassy areas, bushes and shrubs [30]. Indicators of residential buildings, socio-economic activities and tourism are identified using NDBI. Agriculture, plantation and forestry indicators are identified using NDVI. Aquaculture and seaweed aquaculture indicators are identified using NDWI and MNDWI. Indicators of open grass, shrubs and shrubs are identified using SAVI.

The composition and density of land cover are important indicators for assessing the risk of tsunami waves in coastal areas. Classifying and labeling the tsunami hazard level using the SVM, k-NN, RF, ANN or MARS algorithm on VI data will produce land cover predictions for the next season. The spatial distribution of land cover data can be predicted using a computational process using the ordinary kriging method.

The color scale on the spatial object prediction map for areas that are not observation areas shows a different range of values and is interpreted as a tsunami wave hazard level. Red to yellow colors indicate areas of open lands, built-up lands, non-vegetated lands, grasses, bushes and shrubs. Green to blue colors indicate areas of agriculture, plantation and forest or areas of aquaculture ponds, and densely vegetated areas.

The prediction results were tested for their accuracy and validation using the mean square error (MSE) and mean absolute error (MAE) methods. The MSE and MAE accuracy tests are used to calculate the difference between the predicted value and the observed value. If the difference between the two values is getting closer to zero, the prediction results are more accurate.

In the classification process, SVM divides the data set for each vegetation index into two parts and limits them with a line called a support vector. This line is referred to as hyperplane, which is a line that separates two parts that are not connected by n dimensions in Euclidean space (Table 2). The maximum distance between the hyperplane and the vegetation index data set in each space is referred to as optimal hyperplane.

Table 2. Classification of new data classes using SVM algorithms

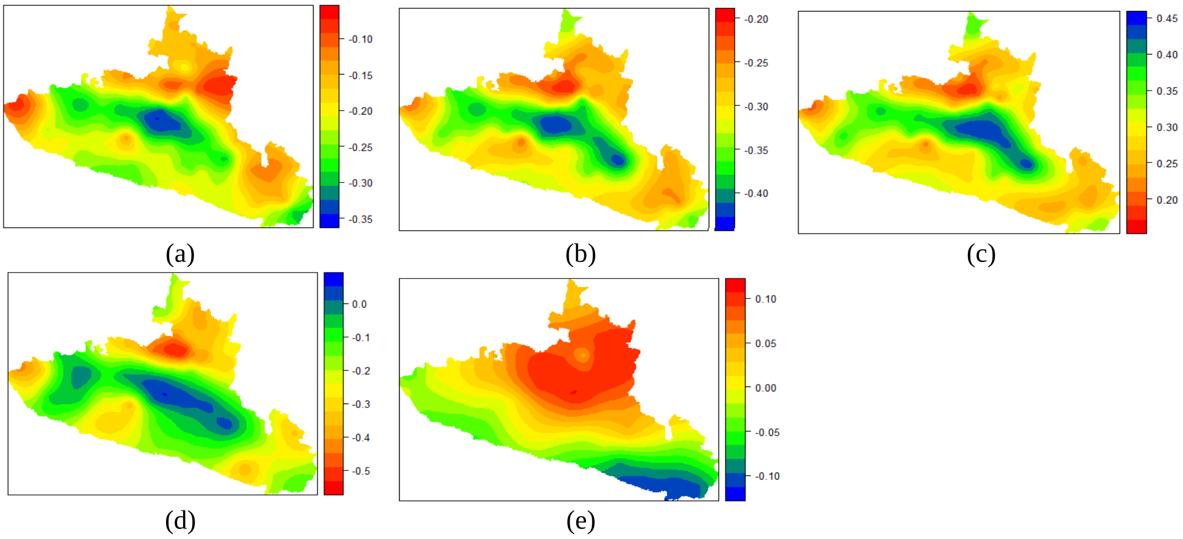
Vegetation Index	Very Low	Low	High	Very High
NDWI	>-0.35-<-0.30	>-0.30 - <-0.25	> -0.25 - < -0.20	>-0.20
MNDWI	>-0.40-<-0.35	>-0.35 - <-0.30	> -0.30 - < -0.25	>-0.25
NDVI	> 0.20-<0.15	>0.15 - <0.10	> 0.10 - < 0.5	>0.5
MSAVI	>-0.50 -<-0.40	>-0.40- <-0.30	>-0.30- <-0.20	>-0.20
NDBI	>-0.10-<-0.05	<-0.05 – 0.00	>0.00-<0.10	>0.20

The SVM algorithm will read each vegetation index data, which data with an extreme value will be labeled as identifier to determine whether the data is a limit on a very low class or a very high class. Each new formed class will be determined by the dividing line of the same two parts, namely the very low-value data class and the very high-value data class as shown in Figure 2.

New classes resulting from SVM classification from historical data are: (a) NDWI, (b) MNDWI, (c) NDVI, (d) MSAVI, and (e) NDBI, which they have different Optimal Hyperplane values, so that they have different data range of width for each class and their interpretations will also be different.

SVM algorithm is a method for recognition, classification and prediction of pattern by forming a multidimensional hyperplane to distinguish between data classes. This algorithm uses a nonlinear kernel function so that it transforms the input space into multidimensional space. SVM algorithm separates data sets into multidimensional forms and creates a hyperplane linear line to optimally separate dimensional data (support vectors). Of the thousands of VI data, there are some data that are not easily separated, so that kernel functions must be formed to form higher dimensions. This data separation results new data of VI which is then tested for statistical accuracy as shown in Table 3. SVM algorithm is accurate in predicting and generating new data on NDBI data with MSE value of 0.003, NDVI and NDWI with MSE value of 0.004.

The k-NN algorithm forms the assumption that the data is not independent and will tend to be near the same or similar data. Formation of data classes in k-NN uses the concept of majority vote, which same or similar data will be in the same class. In vegetation index data, there are 4 classes, namely very low, low, high and very high classes of data set. The criterion used to enter data into classes is the ability to measure the Eucledian distance from the center of data set or data class. The new data class is the result of classification using k-NN algorithms on the vegetation index data as shown in Table 4.



(a) (b) (c)

(d) (e)

Figure 2. New class of SVM historical data classification: (a) NDWI, (b) MNDWI, (c) NDVI, (d) MSAVI, and (e) NDBI which are interpolated with ordinary kriging

Table 3. The prediction test of VI validation using SVM method

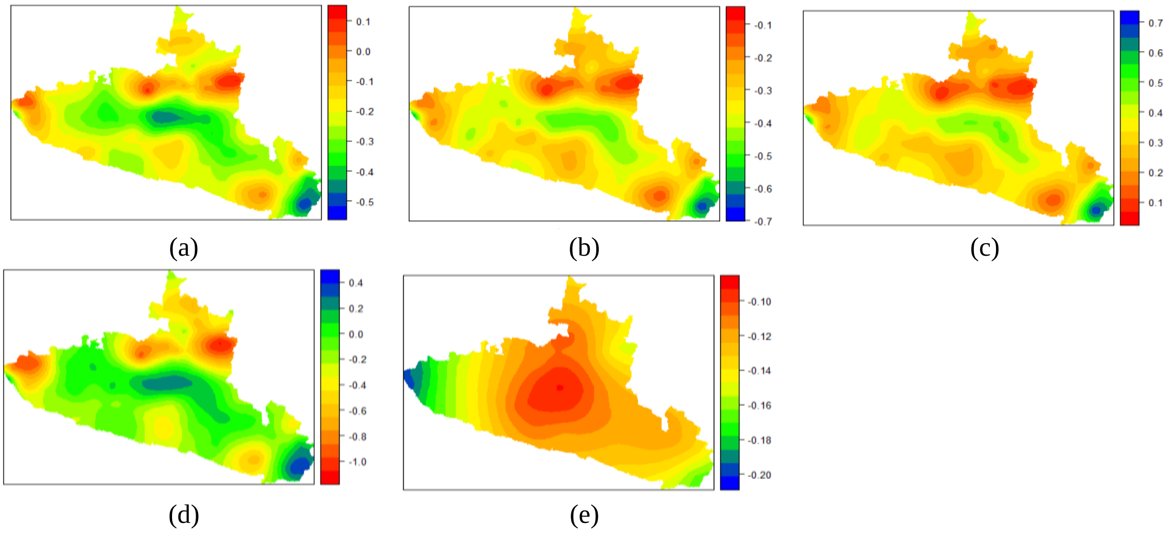
	NDVI	MSAVI	NDWI	MNDWI	NDBI
MSE	0.005	0.018	0.004	0.004	0.003
MAE	0.048	0.084	0.043	0.043	0.011

Table 4. Classification of new data classes using the k-NN algorithm

Vegetation index	Very low	Low	High	Very high
NDWI	<-0.50-<-0.40	>-0.40-<-0.30	>-0.30-<-0.20	>-0.20
MNDWI	<-0.70.-<-0.60	>-0.60-<-0.50	>-0.50-<-0.40	>-0.40
NDVI	<0.10-<0.20	0.20-<0.30	0.3-<0.40	> 0.40
MSAVI	<-1.0	0.20-<0.40	0.40-<0.60	>0.60
NDBI	<-1.0	<-1-<0.80	-0.80-<-0.60	>-0.60

The prediction results using the k-NN algorithms on the vegetation index data are interpolated using ordinary kriging to see the pattern of data distribution as can be seen in Figure 3. The data groups in the new class produced by classification k-NN from historical data are: (a) NDWI, (b) MNDWI, (c) NDVI, (d) MSAVI, and (e) NDBI. In accordance with the concept of k-NN algorithm, the experimental results show that there is a close Euclidean point distance between the VI training data and the VI testing data. The value of the VI classification and prediction results as the new data is determined based on the level of proximity of the VI value of the training data to the testing data (similarity values). The Euclidean VI distance is determined by calculating the square root of the sum of the squared differences between the predicted VI data and the observation point. The closer the Euclidean point distance between the training data and the predicted VI data, the more accurate it is, when measured using the MSE and MAE methods, in which they show values closer to zero (Table 5). The k-NN algorithm is accurate in predicting and generating new data on NDVI, NDWI and MNDWI data with MSE value of 0.001.

The RF algorithm works by forming a large number of decision trees to produce new data that work independently or data that are not correlated with each other. Vegetation index data will be grouped as training data and each group will be arranged as a decision tree. From all decision trees, their average will be calculated as new data which is then classified according to the specified class. The new data classes resulting from classification using the RF algorithm on the vegetation index data are shown in Table 6.



(a) (b) (c)

(d) (e)

Figure 3. New class of k-NN historical data classification: (a) NDWI, (b) MNDWI, (c) NDVI, (d) MSAVI, and (e) NDBI extrapolated with ordinary kriging

Table 5. The prediction test of VI accuracy using the k-NN method

	NDVI	MSAVI	NDWI	MNDWI	NDBI
MSE	0.001	0.003	0.001	0.001	0.073
MAE	0.014	0.024	0.012	0.013	0.003

Table 6. Classification of new data classes using the RF algorithm

Vegetation index	Very low	Low	High	Very high
NDWI	<-0.40-<-0.35	>-0.35-<-0.30	>-0.30-<-0.25	> -0.25
MNDWI	<-0.40-<-0.35	>-0.35-<-0.30	>-0.30-<-0.25	> -0.25
NDVI	<0.20 - <0.25	<0.25-<0.20	<0.20-<0.15	> 0.15
MSAVI	<-0.70 - <-0.60	<-0.60-<-0.50	<-0.50-<-0.40	> -0.40
NDBI	<0.16 - <0.15	<0.15-<0.14	<0.14-<0.13	> 0.13

New classified data are predicted using the RF algorithm using ordinary kriging to see the pattern of data distribution as can be seen in Figure 4. Grouping data into new classes which are not observation areas or not sampling points, spatial distribution predictions are carried out using the ordinary kriging interpolation method as in Figure 4. Groups of data in new classes resulting from classification RF. from historical data are: (a) NDWI, (b) MNDWI, (c) NDVI, (d) MSAVI, and (e) NDBI. RF is an algorithm in the machine learning that works based on the concept of ensemble learning. VI data group is represented by a large amount of data that always increases and becomes more diverse over time. RF randomly forms nodes by selecting the best features and multiple decision trees to produce a more accurate VI prediction value. VI data is in the form of numeric which must be separated based on specific features in the form of classes, which the more various numerical data be separated, the more nodes will be formed and the more decision trees will be generated. The class with the largest number of VI data and the lowest correlation will be used as a predictive VI data model.

The results of the accuracy test of predictive data using RF show a high accuracy as that of other machine learning methods as shown in Table 7. RF algorithm is accurate in predicting and generating new data on NDBI, NDWI and MNDWI data with MSE value of 0.001.

The ANN algorithm works with a large number of nodes that act as processors. The first layer is called the input processor which plays a role in receiving vegetation index data. It is then calculated and the results of the calculation process become input nodes in the next layer. Each node in tier n will be connected to many nodes in tier n-1, the input data is in tier n+1. The new data classes resulting from classification using the ANN algorithm on the vegetation index data are shown in Table 8.

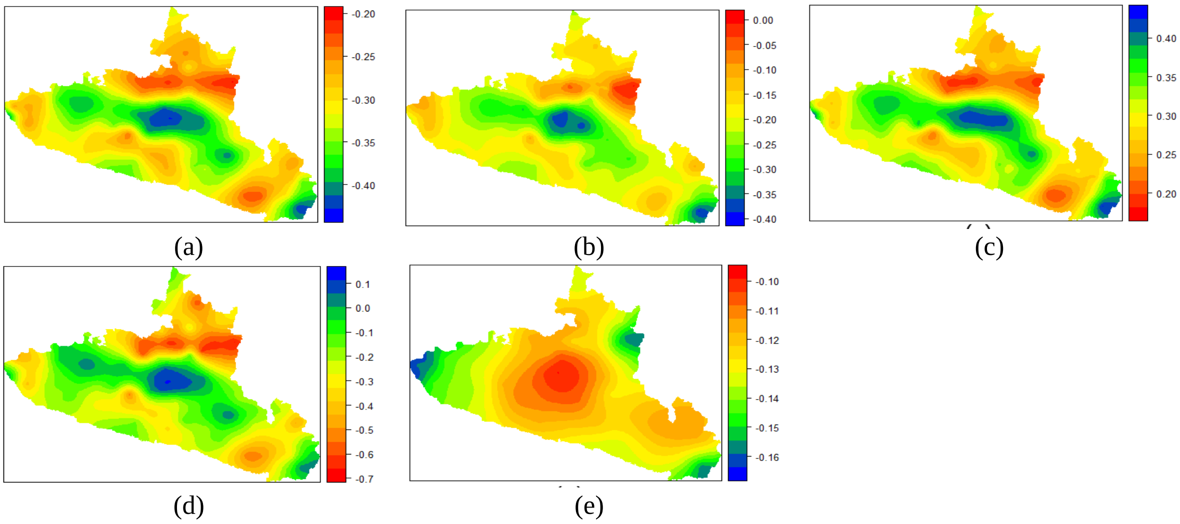


Figure 4. New RF Classification Historical data: (a) NDWI, (b) MNDWI, (c) NDVI, (d) MSAVI, and (e) NDBI interpolated with ordinary kriging

Table 7. Validation test of VI prediction using the RF method

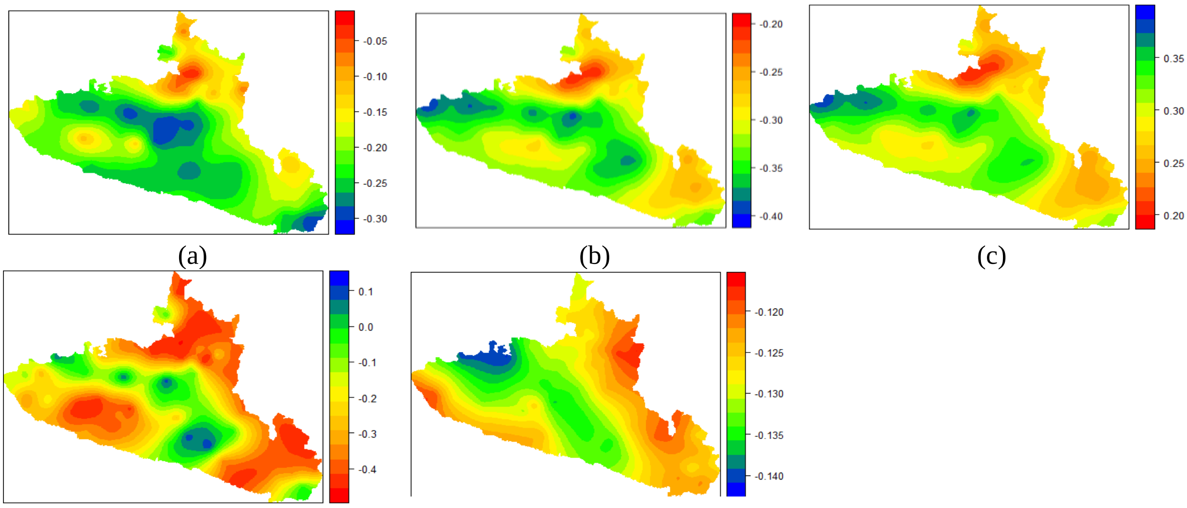
	NDVI	MSAVI	NDWI	MNDWI	NDBI
MSE	0.002	0.006	0.001	0.001	0.001
MAE	0.033	0.055	0.031	0.027	0.007

Table 8. Classification of new data classes using the ANN algorithm

Vegetation Index	Very Low	Low	High	Very High
NDWI	<-0.20- -0.25	-0.25- -0.30	-0.30- -0.35	>-0.35
MNDWI	<-0.05- -0.10	-0.10- -0.15	-0.15- -0.20	>-0.20
NDVI	<0.20 - 0.25	0.25-0.30	0.30-0.35	>0.35
MSAVI	<-0.40 - -0.30	-0.30- -0.20	-0.20- -0.10	>-0.10
NDBI	<-0.140 - -0.135	-0.135- -0.130	-0.130- -0.125	>-0.125

The new data that have been classified using ANN algorithms are predicted using ordinary kriging to see the data distribution pattern as seen in Figure 5. The data is grouped into new classes, and the spatial pattern of distribution is analyzed using the ordinary kriging method as in Figure 5. The grouping of data into new classes resulting from classification using the method ANN from historical data are: (a) NDWI, (b) MNDWI, (c) NDVI, (d) MSAVI, and (e) NDBI. ANN is an algorithm that works based on human brain's system, in the form of a network of neurons and dendrites that transmit external stimuli from sensory organs, transformed into electrical signals that travel along the destination nervous system. VI data enters the input layer node which each node will transform VI data in the hidden layer by adding a weight factor. Each data that has a weight factor is propagated to the next hidden layer node. The results of the prediction data accuracy test using ANN show a high accuracy as other machine learning method (Table 9). ANN algorithm is accurate in predicting and generating new data on NDBI data with MSE value of 0.004, MNDWI data with MSE value of 0.005.

MARS algorithm works by making an assumption that data input and output are linier to create the best classification of vegetation index with a large number of non-linier variables. The new class data from the classification result using MARS algorithm on vegetation index can be seen in Table 10.



(a) (b) (c)

(d) (e)

Figure 5. New Class of Historical Data of ANN Classification: (a) NDWI, (b) MNDWI, (c) NDVI, (d) MSAVI, and (e) NDBI which is interpolated using ordinary kriging

Table 9. Validation test of VI prediction using ANN method

	NDVI	MSAVI	NDWI	MNDWI	NDBI
MSE	0.008	0.024	0.007	0.005	0.004
MAE	0.070	0.116	0.064	0.056	0.016

Table 10. New Class Data from classification result using MARS algorithm

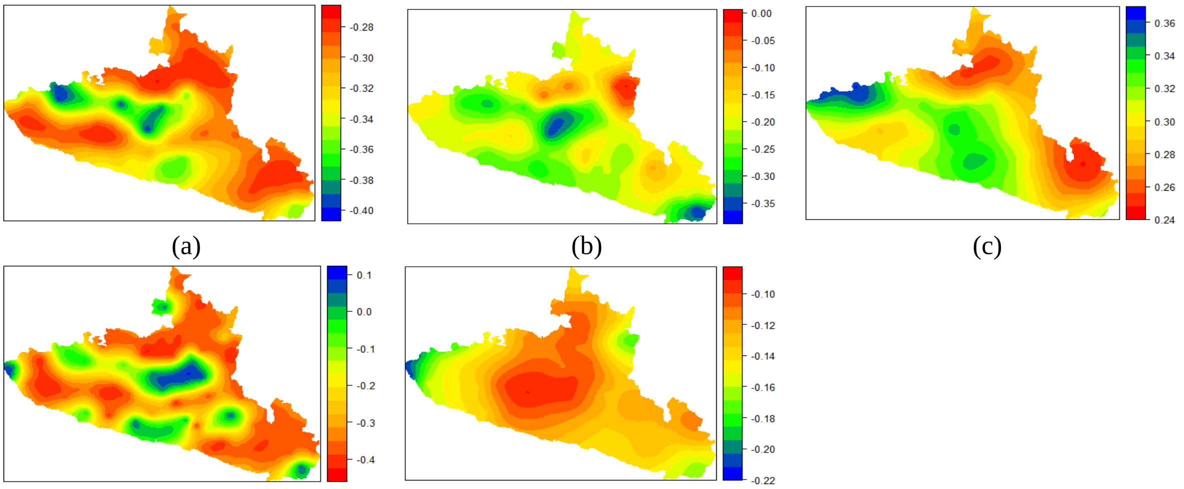
Vegetation Index	Very Low	Low	High	Very High
NDWI	<-0.40- -0.38	-0.38- -0.36	-0.36- -0.34	>-0.34
MNDWI	<-0.35- -0.30	-0.30- -0.25	-0.25- -0.20	>-0.20
NDVI	<0.24 - 0.26	0.26- 0.28	0.28 - 0.30	>0.30
MSAVI	<-0.40 - -0.30	-0.30 - -0.20	-0.20 - -0.10	>-0.10
NDBI	<-0.22 - -0.20	-0.20- -0.18	-0.18- -0.16	>-0.16

New data that have been classified using MARS algorithm are then predicted using ordinary kriging to find out the data distribution pattern as seen in Figure 6. The data is grouped into new classes, and the spatial pattern of distribution is analyzed using the ordinary kriging interpolation method as in Figure 6.

The grouping of data into new classes resulting from classification using the method MARS from historical data are (Figure 6): (a) NDWI, (b) MNDWI, (c) NDVI, (d) MSAVI, and (e) NDBI. The basic idea of MARS algorithm is the existence of non-linear data from polynomial regression on VI data. The algorithm works by assessing each VI data and making knots, then looking for the intersection of the two linear regression line models formed from each VI data point, which the two linear line models will produce new candidate data and hereinafter referred to as the data of the results of VI prediction. In the test of accuracy of the predicted data compared to the observation data of VI, the results of the analysis are shown in Table 11. MARS algorithm is accurate in predicting and generating new data on NDBI data with MSE value of 0.001, and NDVI, NDWI and MNDWI with MSE value of 0.004.

From Table 2, Table 4, Table 6, and Table 8, it can be seen that the most accurate algorithm in predicting new data compared to other algorithms is k-NN as shown in Table 12. k-NN algorithm is accurate in predicting and generating new data on NDBI data with MSE value of 0.004, MNDWI with a value of 0.005 and NDWI with MSE value of 0.007.

On Table 10, it can be seen that k-NN algorithm in this study shows the best performance compared to other algorithms in terms of accuracy results using the MSE and MAE methods. To determine the relief on the earth's surface in the study area, a DEM analyses were performed. DEM shows the relief of the earth's surface by eliminating objects on the ground such as plants and housings, resulting in a smooth surface model. DEM is created using SRTM images to create an elevation model in meters (above the sea level). The experimental results show that the study area, the southern coast of Gunungkidul, Yogyakarta is the zone of hills of Mount Seribu with elevations of 0-400 meters above the sea level, and the field observation show 15 coasts that become the sample of coastal elevation observation for coasts that are located 0-40 meters above the sea level at the distance of 1 to 2 km from the shoreline. The relationship between vulnerability and elevation is shown in Table 13 [30].



(a) (b) (c)

(d) (e)

Figure 6. New Class of MARS Historical Data Classification: (a) NDWI, (b) MNDWI, (c) NDVI, (d) MSAVI, and (e) NDBI extrapolated with ordinary kriging

Table 11. The prediction test of VI validation using the MARS method

	NDVI	MSAVI	NDWI	MNDWI	NDBI
MSE	0.004	0.018	0.004	0.004	0.001
MAE	0.057	0.103	0.052	0.054	0.014

Table 12. The most accurate algorithm on the MSE and MAE accuracy tests

	NDVI	MSAVI	NDWI	MNDWI	NDBI
MSE	k-NN	RF	k-NN	k-NN	MARS
MAE	k-NN	k-NN	k-NN	k-NN	k-NN

Table 13. Relationship between vulnerability and elevation in meters [30]

Elevation (meter)	Vulnerability
5.00 or lower	High Vulnerability
5.00 – 10.00	Rather Vulnerability
10.00 – 15.00	Medium Vulnerability
15.00 – 20.00	Rather Low Vulnerability
20.00 – Higher	Low Vulnerability

Based on visual analysis at the location at the time of observation, calculation of elevation from the SRTM image and the results of the best classification (k-NN algorithm), a new class matrix is developed to determine the level of tsunami hazard on the built index. The new class of tsunami wave hazard level consists of 4 scales and is symbolized by color, namely: Very Low (Blue), Low (Green), High (Yellow) and Very High (Red). Visual analysis includes 6 categories of built-up land objects. The decision matrix and scale for detecting the tsunami wave hazard to built-up land can be seen in Table 14.

The purpose at this stage is to test the accuracy of the class of vegetation indices and land characteristics by comparing observation data and data of the classification of vegetation indices. The test show that the User Accuracy value (observation in the field by researchers) is 91.11% on average, the Producer Accuracy value (interpretation of classification data of vegetation indices from the remote sensing imagery) is 91.51% on average and the Overall Accuracy is 91.12% on average as shown in Table 15.

The Confusion Matrix test is carried out in 6 observational classes, namely: (i) Vegetation Class consisting of forest, shrub and meadow vegetation, (ii) Agriculture Land Class consisting of cultivated vegetation such as agriculture and plantation as indicated by the NDVI value. (iii) Man-Made Structure Class consisting of built-up lands, residential buildings, social facilities and local economic activities of the community as indicated by the NDBI value. (iv) Open Land Class consisting of open land for seasonal crops, bushland, grassland, vegetation and mixed land use as indicated by the MSAVI value. (v) Open Water Class consisting of rivers, aquaculture and rice fields as indicated by the MNDWI value. (vi) Public Mobility Class consisting social and economic activities of the population as indicated by the values of NDBI, NDVI, and MNDWI. Each Class is determined by its coordinate position on the map and matched whether it is visually the same as that of the location. The results of the comparison between the user accuracy and the producer accuracy show that the results of the classification and prediction on the map are the same as those in the field (Table 15).

Table 14. The decision matrix and scale to detect tsunami wave hazard on a built land

Visual analysis	Very low	Low	High	Very high
Symbol	Blue	Green	Yellow	Red
Activities of coastal, swamp, river or estuary fisheries	<-0.50-<-0.40	>-0.40-<-0.30	>-0.30-<-0.20	>-0.20
Activities of irrigated paddy field, agricultural land, or other water bodies	<-0.70.-<-0.60	>-0.60-<-0.50	>-0.50-<-0.40	>-0.40
Activities of agriculture, plantation, forest and vegetation density	<0.10-<0.20	0.20-<0.30	0.3-<0.40	> 0.40
Open land, sandy beach, coastal water vegetation	<-1.0	0.20-<0.40	0.40-<0.60	>0.60
Social and economic activities and built land density in the form of public infrastructure	<-1.0	<-1-<0.80	-0.80-<-0.60	>-0.60
Above sea level elevation (m) [30]	> 15.00	> 10.00 - < 15.00	> 5.00 - < 10.00	< 5.00

Table 15. Testing of the confusion matrix of the class of vegetation indices between the interpretation of vegetation indices prediction using k-NN and the interpretation of vegetation indices classification during the observation at the study area

No	Class	User Accuracy (%)	Producer Accuracy (%)
1	Vegetation (Coastal Forest, Vegetation Land) (NDVI)	93.33	93.33
2	Agriculture Land (Cropland, Horticultural Land) (NDVI)	80.00	92.30
3	Man-made Structure (Bult-up land) (NDBI)	86.66	92.85
4	Open Land (Range Land, Herbaceous and Mixed Rangeland) (MSAVI)	93.33	100
5	Water Surfaces and Water Bodies (MNDWI)	93.33	82.35
6	Public Mobilities (Social and Economic Activity) (NDBI), (NDVI), (MNDWI)	100	88.23
	Overall Accuracy	91.00 %	

4. CONCLUSION

The results show that Sentinel 2A image has a built-up land component and can be extracted using VI algorithms of NDWI, MNDWI, NDVI, MSAVI, and NDBI. Each VI can be classified in a new class using ML, namely SVM, k-NN, RF, ANN, and MARS. Each VI is given a tsunami wave hazard label with a scale of Very Low, Low, High and Very High. Experiments show that the most accurate classification is the k-NN algorithm. The results of the accuracy test using the MSE method are 0.001 of NDWI, 0.003 of MNDWI, 0.001 of NDVI, 0.001 of MSAVI and 0.073 of NDBI. The MAE accuracy test results are 0.014 of NDWI, 0.024 of MNDWI, 0.012 of NDVI, 0.013 of MSAVI and 0.003 of NDBI. The SRTM extraction and analysis shows that the research area at the furthest distance of 1 Km from the coastline has an elevation between 0–15 M asl so that it is in the High Vulnerability to Medium Vulnerability category. The decision matrix for the tsunami hazard resulting from this research has been tested in the field using the Confusion Matrix method with a user accuracy of 91.11%, producer accuracy of 92.16% and average overall accuracy of 91%. This research can be developed by adding the variables of daily mean and maximum height of sea wave and population density within a radius of 2 km from the coastline.

ACKNOWLEDGEMENTS

The authors would like to thank Education and Culture Ministry Republic Indonesia for Grant Research Schema Fundametal Research Regular (PFR), Contract Number No. 182/E5/PG.02.00.PL/2023, Date June, 19 2023 and No. 001/LL6/PB/AL.04/2023, Date June 20, 2023

REFERENCES

[1] M. Meadows and M. Wilson, "A comparison of machine learning approaches to improve free topography data for flood modelling," *Remote Sensing*, vol. 13, no. 2, pp. 1–28, Jan. 2021, doi: 10.3390/rs13020275.

[2] S. Pal, D. Kumar Mishra, A. Haldorai, L. Rama Parvathy, S. Janupriya, and Dv. Babu, "Machine Learning Based Real Time-Heuristic Sensor Data Analytics For Early Warning Prediction," 2021, doi: 10.21203/rs.3.rs-1012679/v1.

[3] L. Yang and G. Cervone, "Analysis of remote sensing imagery for disaster assessment using deep learning: a case study of flooding event," *Soft Computing*, vol. 23, no. 24, pp. 13393–13408, Dec. 2019, doi: 10.1007/s00500-019-03878-8.

[4] A. Ghorbanian, S. A. Ahmadi, M. Amani, A. Mohammadzadeh, and S. Jamali, "Application of artificial neural networks for mangrove mapping using multi-temporal and multi-source remote sensing imagery," *Water (Switzerland)*, vol. 14, no. 2, Jan. 2022, doi: 10.3390/w14020244.

[5] F. Ofli *et al.*, "Combining Human Computing and Machine Learning to Make Sense of Big (Aerial) Data for Disaster Response," *Big Data*, vol. 4, no. 1, pp. 47–59, Mar. 2016, doi: 10.1089/big.2014.0064.

[6] A. E. Maxwell, T. A. Warner, and F. Fang, "Implementation of machine-learning classification in remote sensing: An applied review," *International Journal of Remote Sensing*, vol. 39, no. 9, pp. 2784–2817, 2018, doi: 10.1080/01431161.2018.1433343.

[7] A. Jamali, "Evaluation and comparison of eight machine learning models in land use/land cover mapping using Landsat 8 OLI: a case study of the northern region of Iran," *SN Applied Sciences*, vol. 1, no. 11, pp. 1–11, 2019, doi: 10.1007/s42452-019-1527-8.

[8] M. Panagiota, P. Erwan, G. Philippe, and C. Jocelyn, "Seismic vulnerability assessment using Support Vector Machine classification for remote sensing and in-situ Data."

[9] S. Yulianto, J. Prasetyo, B. H. Simanjuntak, K. D. Hartomo, and W. Sulistyo, "Computer model for tsunami vulnerability using sentinel 2A and SRTM images optimized by machine learning," vol. 10, no. 5, pp. 2821–2835, 2021, doi: 10.11591/eei.v10i5.3100.

[10] S. Y. J. Prasetyo, W. Sulistyo, P. N. Basuki, K. D. Hartomo, and B. Hasiholan, "Computer model of Tsunami vulnerability using machine learning and multispectral satellite imagery," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 2, pp. 986–997, Apr. 2022, doi: 10.11591/eei.v11i2.3372.

[11] L. Moya, C. Geis, M. Hashimoto, E. Mas, S. Koshimura, and G. Strunz, "Disaster Intensity-Based Selection of Training Samples for Remote Sensing Building Damage Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 10, pp. 8288–8304, Oct. 2021, doi: 10.1109/TGRS.2020.3046004.

[12] I. E. Mulia, N. Ueda, T. Miyoshi, A. R. Gusman, and K. Satake, "Machine learning-based tsunami inundation prediction derived from offshore observations," *Nature Communications*, vol. 13, no. 1, 2022, doi: 10.1038/s41467-022-33253-5.

[13] H. Römer *et al.*, "Potential of remote sensing techniques for tsunami hazard and vulnerability analysis-a case study from Phang-Nga province, Thailand," *Natural Hazards and Earth System Science*, vol. 12, no. 6, pp. 2103–2126, 2012, doi: 10.5194/nhess-12-2103-2012.

[14] A. Jamali, "Land use land cover mapping using advanced machine learning classifiers," *Ekologia Bratislava*, vol. 40, no. 3, pp. 286–300, Sep. 2021, doi: 10.2478/eko-2021-0031.

[15] J. Park, Y. Lee, and J. Lee, "Assessment of machine learning algorithms for land cover classification using remotely sensed data," *Sensors and Materials*, vol. 33, no. 11, pp. 3885–3902, 2021, doi: 10.18494/SAM.2021.3612.

[16] M. T. Elnabwy, E. Elbeltagi, M. M. El Banna, M. M. Y. Elshikh, I. Motawa, and M. R. Kaloop, "An approach based on landsat images for shoreline monitoring to support integrated coastal management - A case study, ezbet elborg, nile delta, Egypt," *ISPRS International Journal of Geo-Information*, vol. 9, no. 4, 2020, doi: 10.3390/ijgi9040199.

[17] S. Lee, J. C. Kim, H. S. Jung, M. J. Lee, and S. Lee, "Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea," *Geomatics, Natural Hazards and Risk*, vol. 8, no. 2, pp. 1185–1203, Dec. 2017, doi: 10.1080/19475705.2017.1308971.

[18] K. Nam and F. Wang, "An extreme rainfall-induced landslide susceptibility assessment using autoencoder combined with random forest in Shimane Prefecture, Japan," *Geoenvironmental Disasters*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40677-020-0143-7.

[19] E. Xi, "Image Classification and Recognition Based on Deep Learning and Random Forest Algorithm," *Wireless Communications and Mobile Computing*, vol. 2022, 2022, doi: 10.1155/2022/2013181.

[20] K. Andrej, J. Bešter, and A. Kos, "Introduction to the Artificial Neural Networks, In: Suzuki K (ed), Artificial Neural Networks: Methodological Advances and Biomedical Applications," *InTech*, pp. 1–18, 2011.

[21] A. D. P. Pacheco, J. A. D. S. Junior, A. M. Ruiz-Armenteros, and R. F. F. Henriques, "Assessment of k-nearest neighbor and random forest classifiers for mapping forest fire areas in central portugal using landsat-8, sentinel-2, and terra imagery," *Remote Sensing*, vol. 13, no. 7, Apr. 2021, doi: 10.3390/rs13071345.

[22] Y. Guo, S. Han, Y. Li, C. Zhang, and Y. Bai, "K-Nearest Neighbor combined with guided filter for hyperspectral image classification," in *Procedia Computer Science*, Elsevier B.V., 2018, pp. 159–165. doi: 10.1016/j.procs.2018.03.066.

[23] R. E. McRoberts, "A two-step nearest neighbors algorithm using satellite imagery for predicting forest structure within species composition classes," *Remote Sensing of Environment*, vol. 113, no. 3, pp. 532–545, Mar. 2009, doi: 10.1016/j.rse.2008.10.001.

[24] S. Kuter, G. W. Weber, Z. Akyürek, and A. Özmen, "Inversion of top of atmospheric reflectance values by conic multivariate adaptive regression splines," *Inverse Problems in Science and Engineering*, vol. 23, no. 4, pp. 651–669, May 2015, doi: 10.1080/17415977.2014.933828.

[25] S. Hussain *et al.*, "Spatiotemporal Variation in Land Use Land Cover in the Response to Local Climate Change Using Multispectral Remote Sensing Data," *Land*, vol. 11, no. 5, May 2022, doi: 10.3390/land11050595.

[26] T. A. Akbar, Q. K. Hassan, S. Ishaq, M. Batool, H. J. Butt, and H. Jabbar, "Investigative spatial distribution and modelling of existing and future urban land changes and its impact on urbanization and economy," *Remote Sensing*, vol. 11, no. 2, 2019, doi:

10.3390/rs11020105.

[27] Y. Du, Y. Zhang, F. Ling, Q. Wang, W. Li, and X. Li, “Water bodies’ mapping from Sentinel-2 imagery with Modified Normalized Difference Water Index at 10-m spatial resolution produced by sharpening the swir band,” *Remote Sensing*, vol. 8, no. 4, 2016, doi: 10.3390/rs8040354.

[28] S. N. Tin and W. Muttitanon, “Analysis of enhanced built-up and bare land index (Ebbi) in the urban area of yangon, myanmar,” *International Journal of Geoinformatics*, vol. 17, no. 4, pp. 85–96, Aug. 2021, doi: 10.52939/IJG.V17I4.1957.

[29] L. Halounová, “Reclamation areas and their development studied by vegetation indices,” *International Journal of Digital Earth*, vol. 1, no. 1, pp. 155–164, 2008, doi: 10.1080/17538940701782627.

[30] T. P. T. Sinaga, A. Nugroho, Y. W. Lee, and Y. Suh, “GIS mapping of tsunami vulnerability: Case study of the Jembrana regency in Bali, Indonesia,” *KSCE Journal of Civil Engineering*, vol. 15, no. 3, pp. 537–543, Mar. 2011, doi: 10.1007/s12205-011-0741-8.

BIOGRAPHIES OF AUTHORS

Sri Yulianto Joko Prasetyo completed his doctorate degree on the Doctorate Program of Computer Science, Science Faculty of Gadjah Mada University in 2013. He has been active on research since 2008 until now on the Spatial Data Processing and Remote Sensing. He has published his papers on international journals Scopus Indexed in SIJR 0.8. His main area of interest focuses on geospatial computing. His area of expertise includes machine learning, remote sensing, spatial modeling, and software engineering. His founder of Qua-edutehcno which is a technology based start up for software products higher education quality manajement and technology. He can be contacted at email: sri.yulianto@uksw.edu.

Wiwin Sulistyo completed his study on the Doctorate Program of Computer Science, Science Faculty of Gadjah Mada University Yogyakarta in 2019. He has been active on research since 2018 until now on geography information system. He has published his papers on international journals. His main area of interest focuses on geospatial computing. His area of expertise includes machine learning, remote sensing, spatial modeling, and computer network management. He can be contacted at email: wiwinsulistyo@uksw.edu.

Erwien Christianto is currently working as Lecturer and Researcher at Department of Informatics, Faculty of Information Technology, Satya Wacana Christian University, Salatiga, Indonesia. She has completed his Master in Information Science from Satya Wacana Christian University, Indonesia. His main area of interest focuses network security and IT infrastructure. He can be contacted at erwien.christianto@uksw.edu.

Bistok Hasiholan Simanjuntak is currently working as Lecturer and Researcher at Department of Agrotechnology, Faculty of Agriculture, Satya Wacana Christian University, Salatiga, Indonesia. He has completed his Ph.D. in Soil Science from University of Brawijaya, Indonesia. His main area of interest focuses on plant and soil sciences. His area of expertise includes soil management, land evaluation, geographyc information system, and remote sensing, soil organic matter, soil conservation, and organic farming. Bistok has 20 publications in Scopus journals as author/co-author. He can be contacted at bistok@uksw.edu