IJASST
International Journal of
Applied Sciences and Smart Technologies

Volume 07
Issue 01

p-ISSN 2655-8564, e-ISSN 2685-943, pp. 145-158, 2025

Comparison of SVM, K-NN, RF, CART, and GNB Algorithms for Water Bodies Detection Using Sentinel-2 Level-2a Imagery in Nakhon Pathom, Thailand

Ni Putu Nita Nathalia[1*], Gede Andra Rizqy Wijaya [1,] Kadek Yota Ernanda Aryanto[1], Ni Putu Novita Puspa Dewi[1], Putu Hendra Saputra[1], Ni Putu Karisma Dewi[1], Mellisa Damayanti[1], Kadek Losinanda Prawira[1]

[1]*Engineering and Vocational Faculty/ Universitas Pendidikan Ganesha, Jl. Udayana Singaraja-Bali, 81116, Indonesia*
[*]*Corresponding Author: andra.rizqy@student.undiksha.ac.id*

(Received 14-02-2025; Revised 05-03-2025; Accepted 06-03-2025)

Abstract

Satellite imagery is utilized in various fields, one of which is land use and land cover (LULC) analysis. This study aims to classify water bodies using machine learning models such as SVM, K-NN, RF, CART, and GNB. The data source is obtained from the Google Earth Engine (GEE) platform using Sentinel-2 Level-2A satellite imagery, with a dataset of 5,514 data points per year. The Pixel-Based approach is used as the main method for data extraction, while CRISP-DM is applied as a structured methodology for data management. The parameter indices used include the BSI, NDBI, MNDWI, NDVI and AWEIsh. The results of these calculations serve as dataset features for training algorithms in the model development and training process. Each model has its own parameters, making parameter selection crucial in the training process. Model evaluation is conducted using a confusion matrix. Based on confusion matrix analysis, accuracy, precision, recall, and F1-score are calculated. Among the five models, SVM achieves the highest accuracy at 87%, followed by RF and K-NN. This indicates that the SVM model performs better in binary classification. Ground truth analysis is also conducted using the QGIS platform, which visualizes the classification results, with SVM providing the best visualization.

Keywords: CRISP-DM, Machine learning, Pixel-Based, S2L2A, Water bodies classification

1 Introduction

Water is the most essential element in life. Geographically, water refers to an element that shapes the Earth's surface, such as oceans, rivers, lakes, wetlands, snow, ice, and

water vapor[1]. Physiologically, water serves as a source of life on Earth, as all living organisms require water to survive[2]. However, the role of water is not limited to physiological aspects alone. It also plays a crucial role in economic, social, and environmental activities. For instance, water is used in agriculture for crop irrigation, in industry for production processes, and in recreational activities such as swimming or fishing. Therefore, the identification and monitoring of water areas are necessary to ensure optimal water resource management and to prevent the impacts of water-related disasters[3].

The identification and monitoring process can be carried out using data obtained from remote sensing in space, known as satellite imagery. This approach is chosen because image processing results can be measured in real-time at a relatively low cos. In digital image processing, the classification of water and non-water areas can be performed using a pixel-based image classification approach. Pixel-based image classification is one of the most commonly used methods in land use and land cover (LULC) analysis. This method utilizes digital values to identify each pixel, which is classified into predefined categories based on its characteristic values. A study conducted by Dervisoglu in 2020 on the Duden River in Turkey demonstrated that the pixel-based method has both advantages and disadvantages, depending on data characteristics and analysis objectives[4], [5], [6]. This study aims to analyze the capability of the pixel-based method in classifying water and non-water areas.

The development of machine learning (ML) enhances the data classification process, optimizing image processing for more accurate results. Several classification algorithms are commonly used, including Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Gaussian Naïve Bayes (GNB), and Classification and Regression Tree (CART). Studies have shown that RF outperforms GNB, CART, and GBT in machine learning modelling [7], [8], [9], [10]. Several classification algorithms are commonly used, including Random Forest (RF), Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), Gaussian Naïve Bayes (GNB), and Classification and Regression Tree (CART) [11], [12], [13]. [14] states that the use of the RF algorithm in machine learning modeling performs better than the GNB, CART, and GBT algorithms. [15] states that Google Earth Engine (GEE) has been proven to be an effective and fast

method for LULC mapping. This is demonstrated by the average accuracy of the SVM, RF, and CART models, which are 87.99%, 87.81%, and 84.72%, respectively. This study also shows that the SVM model exhibits better accuracy than other models. Therefore, the authors aim to reanalyze ML algorithms such as SVM, RF, k-NN, GNB, and CART in a different case study with a different dataset.

The case study used in this research is the classification of water and non-water areas in Nakhon Pathom, Thailand. Nakhon Pathom has ponds, rivers, and wetlands that serve as water retention areas. This advantage makes Nakhon Pathom the selected area of interest for obtaining the dataset used in the development of the machine learning (ML) model. This study aims to develop and compare the performance of several ML algorithms previously mentioned in classifying water and non-water areas using Sentinel-2 Level-2A satellite imagery data. The evaluation is based on the accuracy values produced by each model. In addition, an analysis is also conducted on the strengths and weaknesses of each model in the classification process.

Based on these objectives, the benefits of this study include creating an ML model capable of classifying water and non-water areas, as well as providing recommendations on ML algorithms to readers based on the analysis of the strengths and weaknesses of each algorithm used in this research. The evaluation process is carried out using a confusion matrix by calculating the values of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) to derive precision, recall, and F1-score values, which are then used to determine the accuracy of the ML model. With this information, this study can provide readers with insights into the appropriate algorithm for similar case studies.

2 Material and Methods

2.1 Study Area

This study focuses on Nakhon Pathom Province, Thailand. Nakhon Pathom Province is located in the Central Region of Thailand, covering an area of approximately 2,168 square kilometres. Geographically, Nakhon Pathom Province lies between latitudes

13°46′00″N to 14°02′00″N and longitudes 100°00′00″E to 100°20′00″E (**Error! Reference source not found.**). The

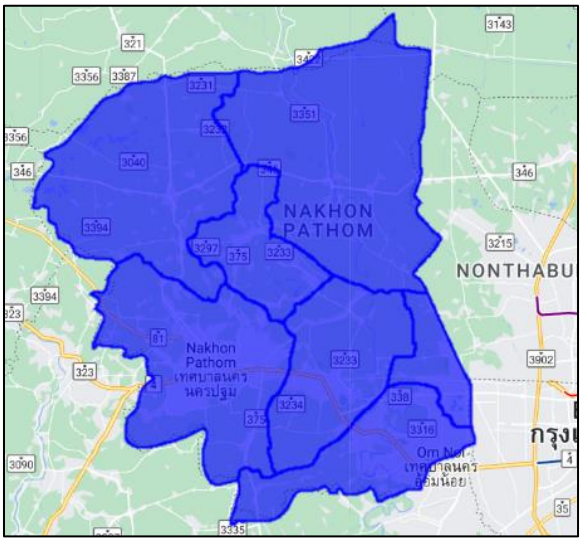

Figure 1. Nakhon Pathom Province, Thailand.

province is situated 56 kilometers west of Bangkok and is administratively divided into seven districts. This province contains water bodies such as rivers, ponds, lakes, and wetlands, making Nakhon Pathom a valuable area of interest for research related to water body classification.

2.2 Cross-Industry Standard Process for Data Mining

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is used as a method to structure a framework for data management in this research (**Error! Reference source not found.**)[16]. The Business Understanding phase has been explained in the introduction section. In this phase, the author analyzes problems that could be caused by water, leading to the idea of forming a machine learning (ML) model capable of classifying water and non-water areas. In the Data Understanding phase, the author explores the necessary data and then collects the data to be processed in the data preparation phase (Table 1).

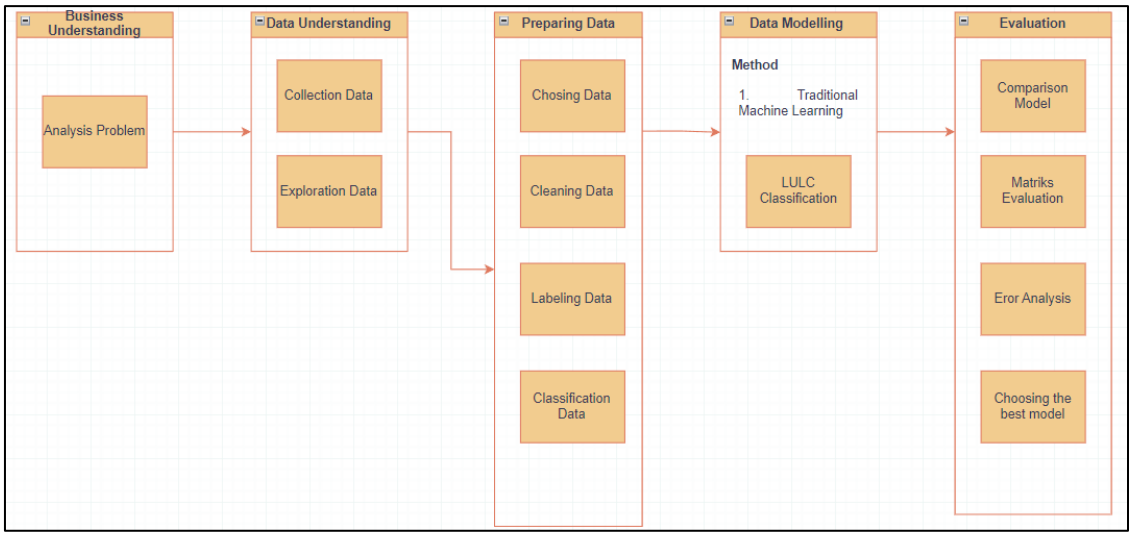

Figure 2. CRISP-DM workflow.

Table 1. Data in a single CSV file

Column	Description
Index	The serial numbers for the data start from 0_0 to 0_299 (for the water class) and 0_0 to 0_99 (for the vegetation and building classes). These numbers will adjust according to the data available.
Spectral Band	The extracted spectral bands include B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B11, and B12.
Class	The data groups are classified as water (1), vegetation (2), and buildings (3).
Geo	The coordinate points correspond to the actual area.

The next phase is Data Preparation. In this phase, the data is cleaned to ensure that the data used is accurate and does not introduce significant bias into the machine learning (ML) model (Table 2. 3). The data is extracted based on classes (i.e., water and non-water classes). Since the data is obtained using Sentinel-2 Level-2A (S2L2A) satellites, the

extraction results consist of the band values available on S2L2A. These values are then calculated using the chosen parameter index, and the results of these calculations are used as dataset features for training the ML model (Table 2. 4). In the modeling phase, the five selected traditional ML algorithms are trained using the processed dataset, with each model having its own parameter tuning (Table 2. 5). The results of this training then move to the Evaluation phase using a confusion matrix (Table 3. 1). In this phase, the best model is selected based on the highest accuracy achieved.

2.3 Satellite Sentinel-2 Level-2A

In S2L2A, the extracted spectral bands consist of 12 bands. The obtained pointing data is then extracted and stored in CSV (comma-separated value) format, which is then processed into a dataset for the learning model (Table 1). This process falls under the category of Data Understanding.

In the Data Preparation phase, data collection was conducted from May 2023 to April 2024. The water data consisted of 300 points per month, vegetation data had 100 points per month, and building data had 100 points per month. The total sample data obtained was 500 points per month, or 6000 points over the course of one year (May 2023 – April 2024) (Table 2. 2). This data will be referred to as the initial data.

Before the extraction phase, we used the built-in feature of S2L2A, namely S2Cloudless, which aims to reduce the impact of clouds so that the values and characteristics of an area can be clearly captured by the satellite. This process is referred to as cloud masking. Therefore, the total sample data obtained after cloud masking is displayed in (Table 2. 3). The final data was then extracted using a Python notebook library within the Visual Studio Code framework. All data extraction was performed through Google Earth Engine (GEE).

Table 2. 2 The total sample points for the initial data

Water	Building	Vegetation	Total
3600	1200	1200	6000

Table 2. 3 Table of final data

Water	Building	Vegetation	Total
3306	1090	1118	5514

The data is then categorized into two classes: water class and non-water class. The band values contained in the downloaded CSV file are then calculated using parameters such as Bare Soil Index (BSI), Normalized Difference Built-up Index (NDBI), Modified Normalized Difference Water Index (MNDWI), Normalized Difference Vegetation Index (NDVI), and Automated Water Extraction Index (AWEIsh). Mathematically, this is written as follows:

$$BSI = \frac{(B11+B4)-(B8+B2)}{(B11+B4)+(B8+B2)} \quad (1)$$

$$NDBI = \frac{(B11-B8)}{(B11+B8)} \quad (2)$$

$$\text{MNDWI} = \frac{(B3-B8)}{(B3+B8)} \quad (3)$$

$$\text{NDVI} = \frac{(B8-B4)}{(B8+B4)} \quad (4)$$

$$\text{AWEIsh} = \frac{(B2+2.5*B3-1.5(B8+B11)-0.25*B12)}{B2+B3+B11+B12} \quad (5)$$

The selection of these five parameters is based on the representation of water, building, and vegetation values in the dataset. BSI and NDBI are used as parameters to help the model recognize buildings, NDVI is used as a parameter to help the model recognize vegetation, and MNDWI and AWEIsh are used as parameters to help the model recognize water. If the value of each parameter is 0.5, this can be interpreted as a situation where the measured characteristic is in the middle of the possible value range, indicating that accurate interpretation may be difficult. For example, an NDVI value of 0.5 may indicate the presence of vegetation in suboptimal conditions or a mix of vegetation and soil, while an MNDWI value of 0.5 may indicate uncertain water presence or areas with high soil moisture. After calculating the parameters, the format of the previous CSV file will change as shown in (Table 2. 4).

Table 2. 4 Dataset for modelling process

Column	Description
Index	The serial numbers for the data start from 0_0 to 0_299 (for the water class) and 0_0 to 0_99 (for the vegetation and building classes). These numbers will adjust according to the data available.
Spectral Bands	The extracted spectral bands include B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B11, and B12.
Class	The data groups are classified as water (1), vegetation (2), and buildings (3).
Geo	The coordinate points correspond to the actual area.
Parameter Index	The values from the calculation of the parameters (BSI, NDBI, MNDWI, AWEIsh, and NDVI).

The next phase is the Modelling phase. The modelling process is carried out using five (5) different ML algorithms, namely SVM, K-NN, RF, CART, and GNB. The steps in the modelling process are almost the same for each model, including defining the x and y variables for the parameters and model class, searching for the best parameters for each model, classification, and performance evaluation. The distinction in the process lies in the selection of parameters (parameter tuning) for each model (Table 2. 5).

Table 2. 5 Parameter tuning for each algorithm

SVM	c: 100; class_weights: None; degree: 2; gamma: ‘scale’; kernel: ‘rbf’.
RF	leaf_size: 20; metric: ‘minkowski’; n_jobs: -1; n_neighbors: 20; p: 2; weights: ‘distance’.

k-NN	bootstrap: True; max_depth: None; min_sampes_leaf: 1; min_samples_spit: 5; n_estimators: 200; n_jobs: -1.
CART	criterion: ‘entropy’; max_depth: 10; min_samles_leaf: 10; min_samples_split: 10.
GNB	var_smoothing: 1e-09.

The final process is evaluation. The evaluation of the model is performed using evaluation matrices such as the confusion matrix. Analysis is carried out on the results based on the accuracy, precision, recall, and F1-score values of each algorithm. Below is the mathematical formulation for calculating accuracy, precision, recall, and F1-score.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

$$\text{F1} - \text{Score} = \frac{2*\text{precision}*\text{recall}}{\text{precision+recall}} \quad (4)$$

Description : TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative).

In addition to the calculations above, ground truth is also used as a form of evaluation and comparison between the model results and the real-world conditions. Ground truthing is performed using the QuantumGIS (QGIS) software.

3 Results and Discussions

3.1 Model Analysis

Based on the results of the confusion matrix, with a dataset of 5514 data points, the model using the SVM algorithm successfully identified 2922 water class samples correctly as water. The model also correctly identified 1870 non-water class samples as non-water. On the other hand, the model incorrectly identified 384 non-water class samples as water, and 338 water class samples as non-water (Table 3. 1). The resulting ratio scale is 6.64:1, or approximately 7:1. Based on the values derived from the confusion

matrix analysis, the results shown in (Table 3. 2) indicate that the SVM model excels in classifying both water and non-water classes with a ratio of 6.64:1 and an accuracy of 87%, followed by the RF and k-NN models with an accuracy of 86%.

Table 3. 1 The acquisition of the confusion matrix values for each model.

	TP	TN	FP	FN	Ratio
SVM	2922	1870	384	338	6.64:1
k-NN	2916	1804	390	404	5.94:1
RF	2918	1822	388	386	6.12:1
CART	2769	1740	537	468	4.48:1
GNB	2663	1629	643	579	3.51:1

Table 3. 2 The calculation of the precision, recall, F1-Score, and accuracy values produced by the model.

Class	*Precision*	*Recall*	*F1-Score*	*Accuracy Model*
		SVM Model		
Water	0.90	0.88	0.89	0.87
Non-water	0.83	0.85	0.84	
		k-NN Model		
Water	0.88	0.88	0.88	0.86
Non-water	0.82	0.82	0.82	
		RF Model		
Water	0.88	0.88	0.88	0.86
Non-water	0.82	0.83	0.82	
		CART Model		
Water	0.86	0.84	0.85	0.82
Non-water	0.76	0.79	0.78	
		GNB Model		
Water	0.82	0.81	0.81	0.78

Non-water	0.72	0.74	0.73

3.2 Visualization of the results of each model using Quantum GIS (QGIS).

The comparison of the results from each model is done for a region in Nakhon Pathom. The blue colour indicates water areas, and the green colour represents non-water areas. Using a pixel-based method, it shows detailed results, but when viewed from a distance, it displays patches, indicating that adjacent areas have been classified into different classes.

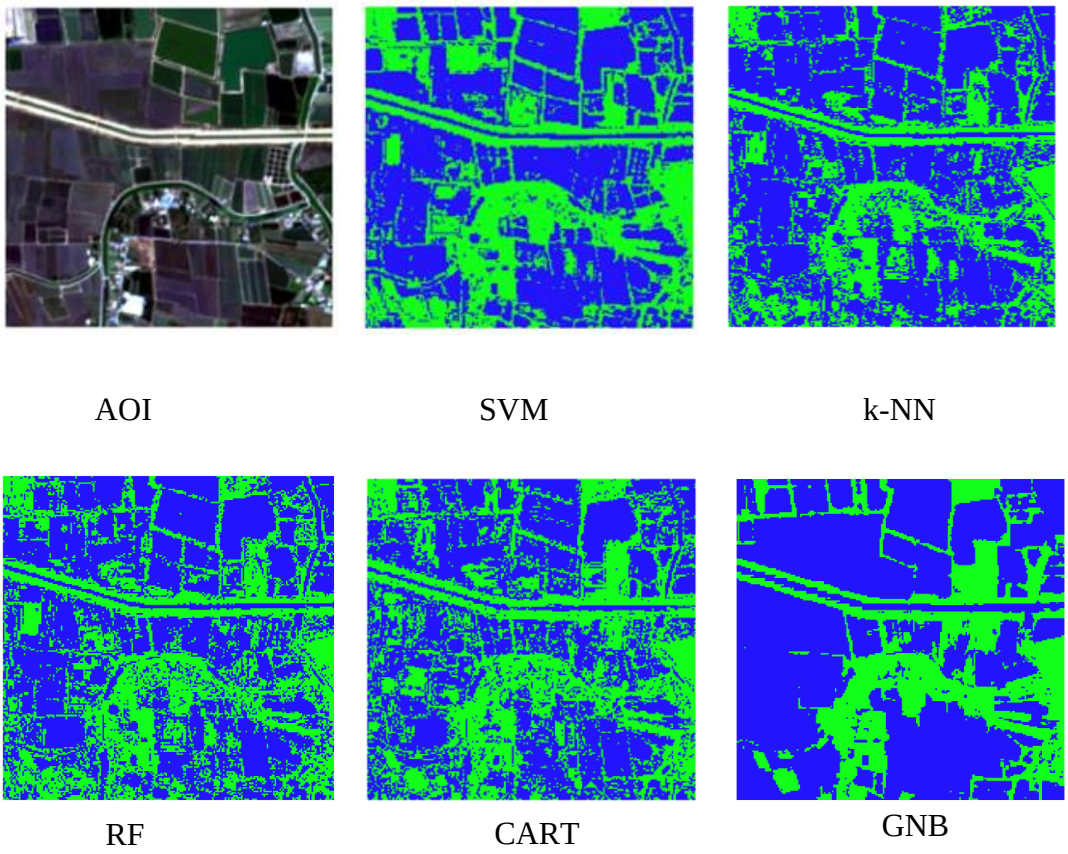

Figure 3. 1 Comparison of the classification results from each model.

The results show that, in the classification of water and non-water areas, the SVM model performs the best, followed by the RF and k-NN models. This can be confirmed based on the accuracy values Table 3. 1 - Table 3. 2 and the provided ground truth results. Therefore, SVM, RF, and k-NN are recommended algorithms for the classification of water and non-water areas.

4 Conclusions

In this study, we understand that the selection of data, algorithms, and methods is crucial to the success of building a machine learning (ML) model. For the classification of water and non-water areas, we recommend several supervised learning algorithms such as SVM, RF, and k-NN. The accuracies achieved by each model are 87%, 86%, and 86%, respectively. To improve the model's quality, it is necessary to increase the dataset size and split the non-water class into more specific classes, such as vegetation, barren land, and buildings. This would result in greater area variability. Future research could consider using deep learning methods like Convolutional Neural Networks (CNNs), which are better at handling the complexity of image data. Additionally, testing the model on datasets from different regions or in varying environmental conditions could provide further insights into the model's generalization ability.

Acknowledgements

The author would like to express gratitude to the supervisor and colleagues who have contributed and assisted the author throughout the process of this research.

References

[1] H. Ghosh, M. A. Tusher, I. S. Rahat, S. Khasim, and S. N. Mohanty, “Water Quality Assessment Through Predictive Machine Learning,” in *Lecture Notes in Networks and Systems*, Springer Science and Business Media Deutschland GmbH, 2023, pp. 77–88. doi: 10.1007/978-981-99-3177-4_6.

[2] W. Jiang *et al.*, “A new index for identifying water body from sentinel-2 satellite remote sensing imagery,” in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Copernicus GmbH, Aug. 2020, pp. 33–38. doi: 10.5194/isprs-Annals-V-3-2020-33-2020.

[3] A. Tassi and M. Vizzari, "Object-oriented lulc classification in google earth engine combining snic, glcm, and machine learning algorithms," *Remote Sens (Basel)*, vol. 12, no. 22, pp. 1–17, Nov. 2020, doi: 10.3390/rs12223776.

[4] A. Dervisoglu *et al.*, "Comparison of Pixel-Based and Object-Based Classification Methods in Determination of Wetland Coastline," *International Journal of Environment and Geoinformatics (IJEGEO)*, vol. 7, no. 2, pp. 213–219, 2020, doi: 10.30897/ijegeo.

[5] L. Qu, Z. Chen, M. Li, J. Zhi, and H. Wang, "Accuracy improvements to pixel-based and object-based LULC classification with auxiliary datasets from google earth engine," *Remote Sens (Basel)*, vol. 13, no. 3, Feb. 2021, doi: 10.3390/rs13030453.

[6] H. Ouchra, A. Belangour, and A. Erraissi, "A Comparative Study on Pixel-based Classification and Object-Oriented Classification of Satellite Image," *International Journal of Engineering Trends and Technology*, vol. 70, no. 8, pp. 206–215, Aug. 2022, doi: 10.14445/22315381/IJETT-V70I8P221.

[7] D. W. Triscowati and A. W. Wijayanto, "Peluang Dan Tantangan Dalam Pemanfaatan Teknologi Penginderaan Jauh Dan Machine Learning Untuk Prediksi Data Tanaman Pangan Yang Lebih Akurat," *Seminar Nasional Official Statistics*, vol. 2019, no. 1, 2020, doi: 10.34123/semnasoffstat.v2019i1.230.

[8] F. Traoré, S. Palé, A. Zaré, M. K. Traoré, B. Ouédraogo, and J. Bonkoungou, "A Comparative Analysis of Random Forest and Support Vector Machines for Classifying Irrigated Cropping Areas in The Upper-Comoé Basin, Burkina Faso," *Indian J Sci Technol*, vol. 17, no. 8, pp. 713–722, Feb. 2024, doi: 10.17485/IJST/v17i8.78.

[9] D. M. Abdullah and A. M. Abdulazeez, "Machine Learning Applications based on SVM Classification: A Review," *Qubahan Academic Journal*, vol. 3, no. 4, pp. 206–218, Nov. 2023, doi: 10.48161/Issn.2709-8206.

[10] M. C. R. Cordeiro, J.-M. Martinez, and S. Peña-Luque, "Remote Sensing of Environment," vol. 253, p. 112209, 2021, doi: 10.1016/j.rse.2020.112209ï.

[11] C. Avci, M. Budak, N. Yagmur, and F. B. Balcik, "Comparison between random forest and support vector machine algorithms for LULC classification," *International Journal of Engineering and Geosciences*, vol. 8, no. 1, pp. 1–10, Feb. 2023, doi: 10.26833/ijeg.987605.

[12] D. Danuri and M. Mohd Pozi, "Machine Learning Approaches for Fish Pond Water Quality Classification: Random Forest, Gaussian Naive Bayes, and Decision Tree Comparison," European Alliance for Innovation n.o., Feb. 2024. doi: 10.4108/eai.21-9-2023.2342964.

[13] H. R. Bittencourt and R. T. Clarke, "Use of Classification and Regression Trees (CART) to Classify Remotely-Sensed Digital Images," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2003, pp. 3751–3753. doi: 10.1109/igarss.2003.1295258.

[14] S. Bayas, S. Sawant, I. Dhondge, P. Kankal, and A. Joshi, "Land Use Land Cover Classification Using Different ML Algorithms on Sentinel-2 Imagery," in *Lecture Notes in Electrical Engineering*, Springer Science and Business Media Deutschland GmbH, 2022, pp. 761–777. doi: 10.1007/978-981-19-0840-8_59.

[15] E. M. Sellami and H. Rhinane, "A New Approach For Mapping Land Use / Land Cover Using Google Earth Engine: A Comparison Of Composition Images," in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, International Society for Photogrammetry and Remote Sensing, Feb. 2023, pp. 343–349. doi: 10.5194/isprs-archives-XLVIII-4-W6-2022-343-2023.

[16] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.