COMPARATIVE ACCURACIES USING MACHINE LEARNING MODELS FOR MAPPING OF SUGARCANE PLANTATION BASED ON SENTINEL-2A IMAGERY IN KEDIRI AREA, EAST JAVA

Ridson Al Farizal Pulungan[1], Rani Nooraeni[2]
[12]Departemen of Statistic, Politeknik Statistika STIS, Jakarta, Indonesia e-mail: ridsonalfarizal15@gmail.com
Received: 18-01-2023; Revised: 25-12-2023; Approved: 26-01-2024

Abstract. Data collection in smallholder sugarcane plantations is still very sensitive to the subjectivity of informants and data collectors. In the meantime, the problem with data collection on sugarcane plantation companies is a low response rate. This situation can reduce the precision of the estimates that are produced. Consequently, the goal of this research is to recognize sugarcane fields using the machine learning models on Sentinel-2A satellite imagery in Kediri Area that covering Kediri Regency and Kediri Municipality, East Java. Along with developing machine learning algorithms, this research will evaluate how well LightGBM performs when compared to other algorithms, including CART, SVM, Random Forest, and XGBoost. Each model employed hyperparameter tuning with random search and stratified 10-fold cross validation to avoid overfitting. The process of labelling satellite imagery using images from Google Street View, then predictor variables used are NDVI, NDWI, NDBI, EVI, and elevation. The most accurate classification model obtained was LightGBM, with a 98% accuracy and a cohen's kappa of 97.7%. The estimated area of sugarcane plantations in the Kediri Regency and Kediri Municipality in September 2022 is 18,897.6 ha and 571.87 ha.

Keywords: *remote sensing,* CART, SVM RBF *kernel*, SVM *polynomial kernel*, *Random Forest*, XGBoost, LightGBM

1 INTRODUCTION

Sugarcane (*Saccharum officinarum*) is a member of the *Gramineae* family, which includes grasses. The sugar and monosodium glutamate (MSG) industries utilize the water extracted from sugarcane stalks as a raw ingredient. (Syathori & Verona, 2020). Thousands of factory workers and sugarcane farmers depend on the sugarcane and MSG industries for a living. Furthermore, sugar has become a necessity for most Indonesians (Sulaiman *et al.*, 2018).

Indonesia's annual sugar consumption continues to rise (BPS, 2022). However, the sugarcane plantations area did not rise considerably; in fact, it decreased because of land conversion. As part of its attempts to establish national food security, the Indonesian government is attempting to attain self-sufficiency in sugar production (Sulaiman *et al.*, 2018).

Accurate and up to date data on sugarcane plantations is required to progress sugarcane plantations in Indonesia. Statistical Agency (BPS) and the Directorate General of Plantations are the two primary data sources for sugarcane plantations in Indonesia. BPS obtained data related to sugarcane plantation companies using Computer Assisted Web Interviewing (CAWI)-based self-enumeration and by interviewing companies that had not filled out the form. (BPS, 2022).

Meanwhile, the Directorate General of Plantations collects statistics on smallholder plantations through field officers' estimations. Officers will collect data on planting methods, population density per hectare, land area (distinct from planting area), and other factors.

The sources of information included planters, farmer groups, village officials, and others. Officers will estimate the area based on this data in accordance with the Guidelines for Implementing Plantation Commodity Data Management (PDKP) (*Kementrian Pertanian*, 2013).

However, until recently, data gathering on smallholder plantations was very sensitive to informant and data collector subjectivity (Ruslan & Prasetyo, 2021). The guidelines for creating predictions that are not up to date can provide estimates that are not in agreement with the present circumstance. Moreover, the problem with data gathering on sugarcane plantation companies is a low response rate. These factors can reduce the precision of the estimates that are produced.

Estimation results that are either overestimated or underestimated can lead to policymaking errors, particularly in the case of sugar import regulations. When the supply of sugar from local farmers is sufficient, excessive imports of sugar might cause losses to sugar farmers or even go bankrupt. In the meanwhile, when the sugar supply is insufficient, low imports will force the price of sugar and its products to decrease uncontrollably.

Various methods of data gathering can be utilized to improve the quality of sugarcane plantation data, including remote sensing (Ruslan & Prasetyo, 2021). Numerous research on the subject have been conducted, including those by Wang *et al.,* (2020), Cevallos *et al.,* (2019), Jiang *et al.*, (2019), Mulianga *et al.*, (2015), Schultz *et al.*, (2015). Previous research indicates that the extent of sugarcane plants can be accurately recognized using the machine learning model applied to satellite data. Some machine learning methods that have been used previously include CART (Verma *et al.,* 2017; Wang *et al.,* 2020), SVM (Everingham *et al.,* 2007; Wang *et al.,* 2019), Random Forest (Jiang *et al.,* 2019; Schultz *et al.,* 2015), and XGBoost (Jiang *et al.,* 2019).

The CART, Random Forest, and XGBoost methods are tree-based models that do not require a lot of preprocessing data but perform well in detecting sugar cane plantations. (Jiang *et al.*, 2019; Schultz *et al.*, 2015; Wang *et al.,* 2020). Moreover, the SVM method works exceptionally well with high dimensional data (Ghaddar *et al.*, 2018; Som-Ard *et al.*, 2021). Other research demonstrates that the XGBoost algorithm detects sugarcane plantation areas with a similar accuracy to the Random Forest method, but at a considerably faster rate (Som-Ard *et al.*, 2021).

Along with the development of machine learning algorithms, there are now various kinds of machine learning algorithms that are claimed to be more efficient and provide more accurate prediction results. The LightGBM algorithm is claimed to be much more efficient than XGBoost but still preserves the accuracy value (Ke et al., 2017). Research by McCarty et al. (2020) has shown that the LightGBM algorithm is more efficient and performs better than SVM and Random Forest in classifying land use and land cover over large geographic areas. The LightGBM has provided much better performance and efficiency than SVM, Random Forest, and KNN in tree species classification in Portugal with multi-temporal Sentinel-2A data (Łoś et al., 2021).

However, the use of the LightGBM algorithm is still rarely used in classifying sugarcane plantations. Consequently, the goal of this research is to analyze the performance of LightGBM compared to CART, SVM, Random Forest and XGBoost in detecting sugarcane plantations on sentinel-2A satellite images. Furthermore, this research will estimate the area of sugarcane crops in the Kediri region in September 2022 using the best model.

2 MATERIALS AND METHODOLOGY

2.1 Study Area

The focus of this research is in Kediri Area that covering Kediri Regency and Kediri Municipality, East Java Province. This is because East Java Province has the most sugarcane plantations in Indonesia, accounting for approximately 44.09 percent of Indonesia's total sugar production in 2021. Malang Regency and Kediri Regency are the two districts with the most sugarcane production in East Java Province (BPS, 2022). In this research, Kediri Regency was selected because it has a lower cloud cover percentage than Malang Regency. Kediri Municipality was also included in the study area because of its existence in Kediri Regency.

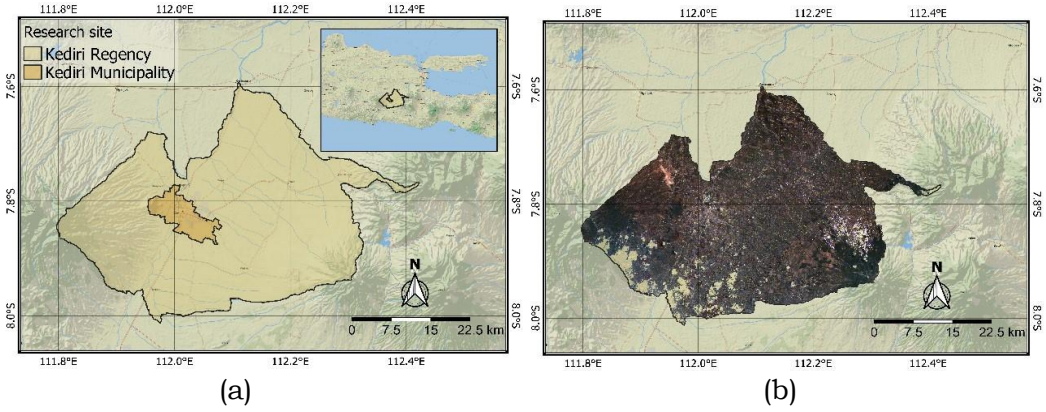


Figure 2-1: (a) Research site, Kediri Area East Java. (b) Sentinel-2A imagery of research site.

2.2 Data Collection

The satellite image used is Sentinel-2A MSI for September 2022 sourced from Google Earth Engine (GEE). The month of September was selected because Google Street View (GSV) shooting in the Kediri Area was primarily conducted in September 2022. In addition, the satellite imagery is incorporated into the cloud masking procedure to eliminate cloud cover.

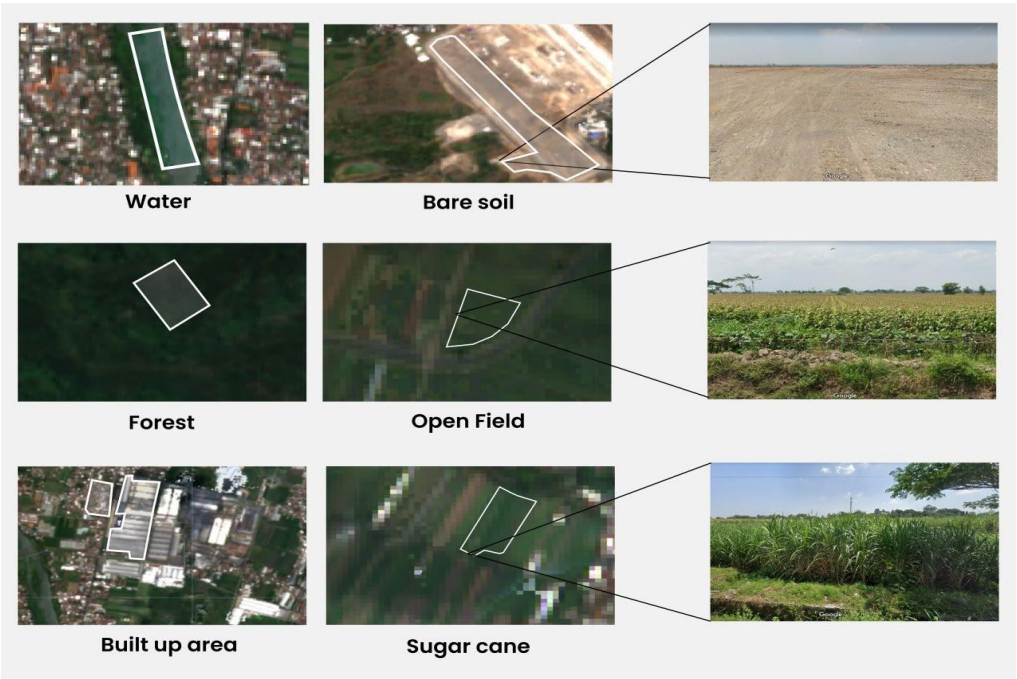


Figure 2-2: The process of labelling Sentinel-2A imagery with GSV

Table 2-1: Number of pixels in sample data.

Class	Class Name	Sample size
1	Water	458
2	Bare soil	455
3	Built up area	465
4	Open field	450
5	Sugar cane	481
6	Forest	495
	Total	2.804

Using GSV to identify satellite imagery. This is owing to the lack of formal administrative maps of sugarcane fields in Indonesia and the restricted scope of direct field surveys. The process of labelling satellite imagery using only images from GSV was carried out in September 2022. The sample polygons were categorized into six groups: water bodies, bare soil (non-vegetative and fallow land), built-up area (roads and buildings), open field (non-sugarcane and non-forest), sugarcane, and forests. Then, points are selected at random from the polygon, with the number of points for each class listed below.

2.2 Feature Collection

Sentinel-2A MSI imagery consists of 13 spectral bands with three different spatial resolutions with ground sampling distances of 10, 20, and 60 meters (Nurmasari & Wijayanto, 2021). The spectral band specifications of the MSI Sentinel-2A instrument can be seen in table 2-2.

Table 2-2: Sentinel-2A MSI instrument spectral band specifications.

Band	Resolution	Central Wavelength	Description
B1	60 m	443 nm	Ultra Blue (Coastal and Aerosol)
B2	10 m	490 nm	Blue
B3	10 m	560 nm	Green
B4	10 m	665 nm	Red
B5	20 m	705 nm	Visible and Near Infrared (VNIR)
B6	20 m	740 nm	Visible and Near Infrared (VNIR)
B7	20 m	783 nm	Visible and Near Infrared (VNIR)
B8	10 m	842 nm	Visible and Near Infrared (VNIR)
B8a	20 m	865 nm	Visible and Near Infrared (VNIR)
B9	60 m	940 nm	Short Wave Infrared (SWIR)
B10	60 m	1375 nm	Short Wave Infrared (SWIR)
B11	20 m	1610 nm	Short Wave Infrared (SWIR)
B12	20 m	2190 nm	Short Wave Infrared (SWIR)

In general, composite indices are derived from these spectral bands; each composite index has a specific application. The *Normalized Difference Vegetation Index* (NDVI) has been utilized extensively to detect sugarcane fields (Cevallos *et al.,* 2019; Jiang *et al.*, 2019; Mulianga *et al.*, 2015; Schultz *et al.*, 2015). In addition, based on research by Mulianga et al., (2015) *Normalized Difference Water Index* (NDWI) gives better classification results than using NDVI. Another research by Nonato and Oliveira (2013) demonstrated that a combination of NDVI and EVI could accurately categorize sugarcane. Additionally, the Normalized Difference Built-Up Index (NDBI) composite index is utilized to differentiate between different bare soil classes and built-up area (Marsuhandi. *et al.*, 2020). The generalised formula for computing each composite index is as follows:

$$NDVI = \frac{NIR - RED}{NIR + RED} \quad (2\text{-}1)$$

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR} \quad (2\text{-}2)$$

$$NDWI = \frac{GREEN - NIR}{GREEN + NIR} \quad (2\text{-}3)$$

$$EVI = 2.5 \times \frac{NIR - RED}{NIR + 6 \times RED - 7.5 \times BLUE + 1} \quad (2\text{-}4)$$

In addition to composite indices, maps of regional elevations are also used to improve accuracy. This is because sugar cane can only grow below an altitude of 1,400 meters above the sea level, and because growth tends to be slow, sugar cane is typically not planted at an altitude of 1,200 meters above sea level. Below 500 meters above sea level is the best land elevation. (Indrawanto et al., 2010). The elevation map used is derived from National Aeronautics and Space Administration (NASA) Shuttle Radar Topography Mission (SRTM) Digital Elevation (DEM) 30 m.

2.3 Methods

The steps of this research are divided into three parts: preprocessing, processing, and estimation. The preprocessing step comprises of filtering satellite images with a cloud percentage of less than 20% and cloud masking of satellite imaging data; this stage is executed using GEE. At the processing stage, the satellite image data and the labels that have been made are extracted and splitting data is carried out with the proportion of training data being 80% and 20% testing data. Stratified random sampling is used to splitting data so that the proportion of each class in the resulting dataset is not significantly altered (Kuhn & Silge, 2022). The utilized machine learning models are Classification and Regression Trees (CART), Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient-Boosting Machine (LightGBM).

2.3.1 CART (Classification and Regression Tree)

CART is a development of the previous decision tree algorithm, namely ID3 and C4.5, where the CART algorithm is able to handle classification and regression cases. CART is a simple method and does not require a long time in determining the best parameters (Tariq et.al. 2023).

The CART algorithm uses the highest gini gain value as a criterion in splitting data. Suppose a dataset S with size $N \times (j+1)$ contain response variable and predictor variables, where $j = 1, \dots, p$ denotes the number of predictor variables. Then the data is split into S_L when $X_j \leq x_{j(i)}$ and the remaining observations to S_R. Then the formula for calculating gini gain can be seen in equation 2-5:

$$Gini\ gain = gini(y) - \frac{N_L}{N} gini(y_L) - \frac{N_R}{N} gini(y_R) \quad (2\text{-}5)$$

Where $gini(y_L)$ is the gini impurity/gini index value of the y variable in S_L, while $gini(y)$ is the gini impurity value of the y variable in S. In the classification case, the metric used in gini gain is the gini impurity/gini index. While in regression, the metric used is variance (Strobl et al., 2007). The formula for calculating gini impurity/gini index and variance can be seen in equations 2-6 and 2-7:

$$Gini\ Impurity = 1 - \sum_{i=1}^{C} p_1^2 \quad (2\text{-}6)$$

$$Variance = \frac{\sum_{i=1}^{n} (y_i - \bar{y})_2}{n-1} \quad (2\text{-}7)$$

Where C is the number of class labels. The process of finding the highest gini gain value will be done for all predictor variables and each class label.

2.3.2 Random Forest

Random forest is a bagging method (Bootstrap Aggregating) in several decision tree algorithms. The result of the most votes will be the prediction output in the classification case, while in the regression case, using the average value of each decision tree prediction. Not all predictor variables nor all observations will be included in every decision tree algorithm. However, it is a sample of observations and a sample of predictor variables (Tariq et.al. 2023). By doing this, the risk of multicollinearity will decrease because the trees naturally do

not correlate through this process. (McCarty et al. 2020). The Random Forest algorithm has the advantage that it does not require a lot of data preprocessing and provides high accuracy in the detection of sugarcane plantations (Jiang et al., 2019; Schultz et al., 2015).

2.3.3 SVM (Support Vector Machine)

SVM is a classification method that creates nonlinear constraints by creating linear constraints within a larger, modified feature space. The SVM algorithm has several kernel functions, among which are linear, polynomial, radial basis function (RBF) and sigmoid. In our research, we use a radial basis function (RBF) and polynomial kernel due to some of the composite indices not being linearly separable (McCarty et al. 2020). The SVM algorithm usually normalizes the predictor variable before it is entered into the model; this is done to get optimal results. (Kuhn & Silge, 2022). Please see Cortes & Vapnik (1995) for a more detailed mathematical explanation of this method.

2.3.4 XGBoost

The XGBoost algorithm introduced by Chen and Guestrin, is a classification algorithm that uses the concept of boosting, which is an iterative process to strengthen weak classifiers so that the longer the classifier becomes the higher the performance. The XGBoost algorithm uses a decision tree to calculate the residual value, which is the difference between the output of the base model classifier and the target value. At the beginning, this residual value will be large because the classifier has not learned the pattern of the target. But as the iteration increases, the XGBoost algorithm will add a decision tree so that the residual value gets smaller. Other research demonstrates that the XGBoost algorithm detects sugarcane plantation areas with a similar accuracy to the Random Forest method, but at a considerably faster rate. (Som-Ard et al., 2021). This is because the XGBoost algorithm supports the use of GPU (Graphics Processing Unit). For more details of XGBoost algorithm, please refer to Chen & Guestrin (2016).

2.3.5 LightGBM

LightGBM (Light Gradient Boosting Machine) is a combination of the Gradient Boosting Decision Tree (GBDT) algorithm with the Exclusive Feature Bundling (EFB) and Gradient-Based One Side Sampling (GOSS) algorithms to handle huge data with preserving accuracy (Ke et al., 2017). The GOSS algorithm aims to eliminate data with small gradients because data with large gradients has a more important role in decision-making. While the EFB algorithm is a technique that aims to reduce the number of features by grouping exclusive features. However, because determining the combination of features in an optimal way is quite difficult, the histogram algorithm on LightGBM is used to make its implementation. LightGBM's mathematical explanation can be seen in more detail in Ke et al., (2017).

2.3.6 Accuracy and Cohen's Kappa Assessment

The SVM algorithm requires normalization of predictor variables to get optimal results (Kuhn & Silge, 2022). In this study, the predictor variables will be normalized first for the SVM model. Each model's parameters are optimized by hyper-parameters tuning and random search. In addition, stratified 10-fold cross validation is employed to prevent overfitting. stratified 10-fold cross validation was selected so that the distribution of each class in each fold would be about the same as the distribution in the initial dataset. (Prusty *et al*, 2022, Pramana *et al.*, 2018). In our research, we used an R library called *tidymodels* to perform the entire workflow from data preprocessing to model evaluation. The final parameters used for all models in this study are as follows:

Table 2-3: Model parameters.

Model	Parameters
LightGBM	1. Maximum depth of trees [1, 15] 2. Minimal node size [2, 40] 3. Number of trees [1, 2000] 4. Minimum loss reduction (transformed scale) $[10^{-100}, \text{inf}]$ 5. Learning rate $[10^{-100}, \text{inf}]$ 6. Iterations before stopping [3, 20]
XGBoost	1. Maximum depth of trees [1, 15] 2. Minimal node size [2, 40] 3. Number of trees [1, 2000] 4. Minimum loss reduction (transformed scale) $[10^{-100}, \text{inf}]$ 5. Learning rate $[10^{-100}, \text{inf}]$ 6. Iterations before stopping [3, 20]
Random Forest	1. Number of trees [1, 2000] 2. Minimal node size [2, 40]
CART	1. Maximum depth of trees [1, 15] 2. Minimal node size [2, 40]
SVM RBF Kernel	1. Cost (transformed scale) [-10, 5] 2. Radial basis function sigma (transformed scale) [-10, 0] 3. Insensitivity margin [0, 0.2]
SVM Polynomial Kernel	1. Cost (transformed scale) [-10, 5] 2. Degree [1, 3] 3. Scale factor (transformed scale) [-10, -1]

The best model is selected using accuracy and cohen's kappa values. Because the number of samples for each class is somewhat similar, the accuracy metric is utilized (Grandini *et al.*, 2020). In contrast, cohen's kappa was chosen because it provides a more accurate evaluation of model performance by taking the number of observations between classes into account (Landis & Koch, 1977). The computation for the accuracy and cohen's kappa was as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2\text{-}5)$$

$$Cohen's\ kappa = \frac{\Pr(a) + \Pr(e)}{1 - \Pr(e)} \quad (2\text{-}6)$$

Where:
TP : true positive
TN : true negative
FP : false positive
FN : false negative
Pr(a) : probability of the correctly classified
Pr(e) : probability of expectation between classes

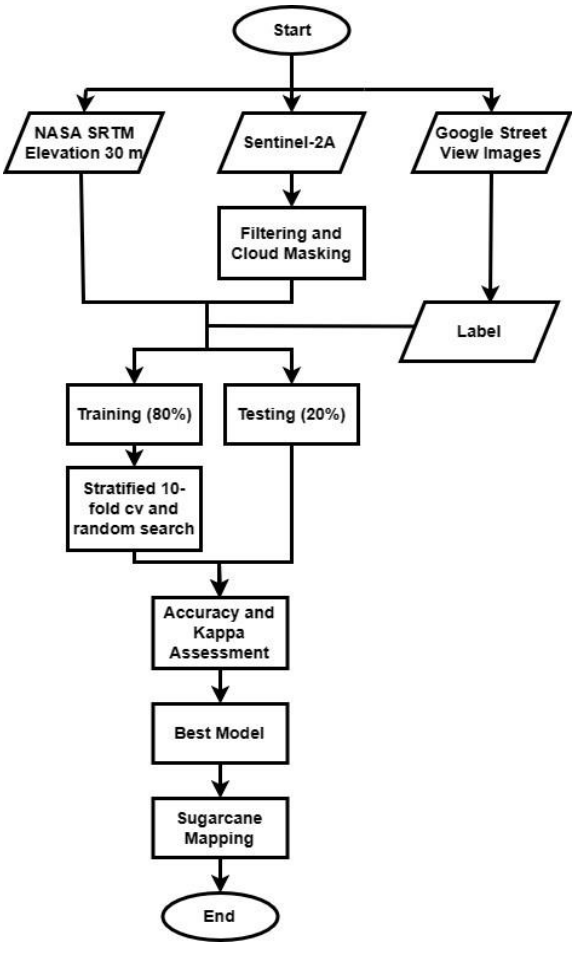


Figure 2-3 depicts the general progression of this research methodology:

During 2021 and September 2022, the area of sugar cane plants in the Kediri Area will be estimated using the best model obtained. The predicted area of

sugarcane planting in 2021 will be compared to official *statistic* data from BPS-published. Using the following formula, the approximate area of sugarcane plantations is determined:

$$A = Ps \times L \qquad (2\text{-}7)$$

where:
A : Sugarcane plantation area
Ps: Percentage of pixels of sugarcane plantations
L : Regency/Municipality area

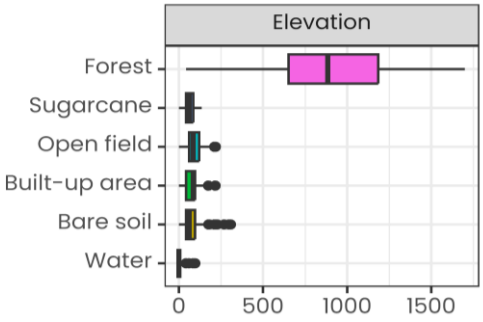
3 RESULTS AND DISCUSSION

3.1 Data Exploration

Figure 3-1 boxplots shows that using the NDVI, NDWI and EVI composite indices, the distinctions between forest classes, open field, and sugarcane plantations can be effectively captured. In contrast, the NDVI, NDWI, and EVI do not capture the distinction between land and built-up land as clearly as the NDBI does. In addition, the variation in the range of elevation values between sugarcane plantations, forest, and open field terrain can be utilized to differentiate the three classifications.

3.2 Classification Result

Hyperparameter tuning has been performed as many as 25 combinations of parameters for each model. Furthermore, each model has been stratified 10-fold cross validation to obtain true values for accuracy and cohen's kappa. Based on the outcomes of hyperparameter tuning with random search, the optimal parameters for every model can be seen in Table 3-1.



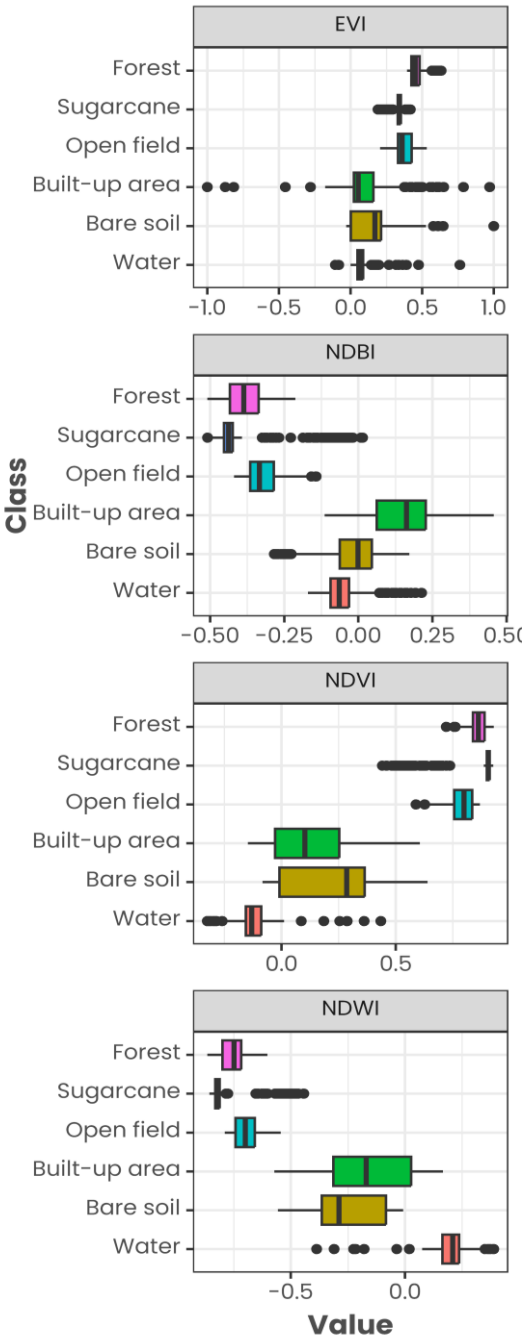


Figure 3-1: Distribution of composite index values per class

The parameter optimized LightGBM model has more accuracy and cohen's kappa value than the other models, as shown in Figure 3-2. Furthermore, the LightGBM model also has a small confident interval for accuracy and cohen's kappa. This indicates that the performance differences across the LightGBM variants are not significant. The same condition also occurs in the

Random Forest, SVM, and CART algorithms.

Table 3-1: Optimal parameters for each model.

Model	Parameters
LightGBM	Maximum depth of trees: 13 Minimal node size: 4 Number of trees: 1423 Minimum loss reduction: 2.04×10^{-9} Learning rate: 3.7×10^{-3} Iterations before stopping: 6
XGBoost	Maximum depth of trees: 12 Minimal node size: 8 Number of trees: 760 Minimum loss reduction: 5.53×10^{-5} Learning rate: 0.151 Iterations before stopping: 5
Random Forest	Number of trees: 693 Minimal node size: 3 Maximum depth of trees: unlimited depth
CART	Maximum depth of trees: 11 Minimal node size: 4
SVM RBF Kernel	Cost: 20.6 Radial basis function sigma: 0.329 Insensitivity margin: 0.196
SVM Polynomial Kernel	Cost: 0.124 Degree: 2 Scale factor: 0.0438

The LightGBM model achieves a maximum accuracy of 97.0%, while kappa is 96.3%. Random Forest, with a 96.2% accuracy and 95.5% kappa, was the second-best model obtained. However, the Random Forest model produces a smaller standard error than other models. In addition, it can be observed that the tree-based model tends to generate superior results compared to other models. Tables 3-2 and 3-3 provide a comparison of each model's precision, Cohen's kappa, and standard error values.

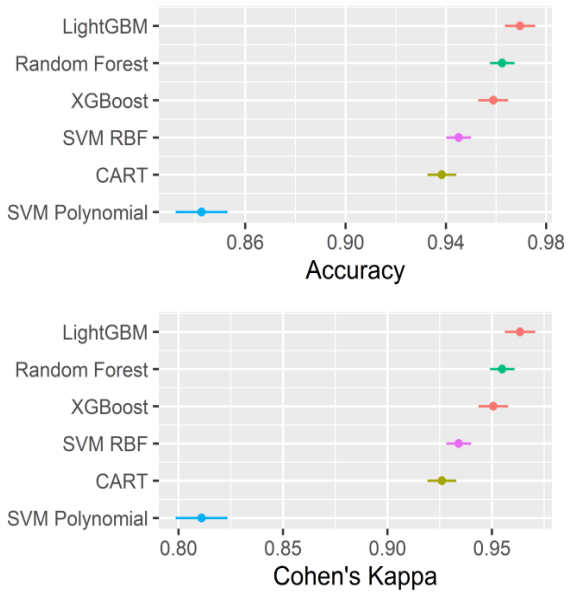


Figure 3-2: Estimated accuracy, cohen's kappa and approximate confidence intervals for the best model.

Table 3-2: Result of accuracy.

Model	Accuracy	Std error
LightGBM	**0.970**	0.00556
Random Forest	0.962	**0.00445**
XGBoost	0.959	0.00549
SVM RBF kernel	0.945	0.00451
CART	0.938	0.00525
SVM polynomial kernel	0.843	0.00990

Table 3-3: Result of cohen's kappa

Model	Cohen's kappa	Std error
LightGBM	**0.963**	0.00668
Random Forest	0.955	**0.00534**
XGBoost	0.951	0.00659
SVM RBF kernel	0.934	0.00541
CART	0.926	0.00630
SVM polynomial kernel	0.811	0.01190

As the final measure of a model's success, each model will be evaluated using test data in the next phase. Table 3-4 demonstrates that the LightGBM model is still the best model with the highest accuracy and kappa, with an accuracy of 98% and a Cohen's kappa of 97.7%. There is no overfitting or

underfitting because the accuracy and Cohen's kappa values between the training data and the testing data are not significantly different, showing that the final model is adequate.

Table 3-4: Model performance comparison in the test data.

Model	Accuracy	Cohen's Kappa
LightGBM	**0.980**	**0.977**
Random Forest	0.970	0.964
XGBoost	0.968	0.962
SVM RBF kernel	0.952	0.942
CART	0.940	0.927
SVM polynomial kernel	0.868	0.842

The next phase is that each model will be used to classify land cover in the Kediri Area. The classification results for each model can be seen in Figure 3-3. Based on Figure 3-3 all models can classify forest classes very well. However, SVM polynomial kernel models tend to classify bare soil and open field classes over other classes. This is likely because the best SVM polynomial models are obtained only with degree 2 (see Table 3-1). Some classes may not be quadratically separated from each other, so the SVM polynomial kernel models cannot be classified properly.

SVM RBF kernel and CART models tend to classify open field classes into built-up area classes. However, LightGBM and Random Forest models are robust models (Zheng et al., 2023). This can be seen in the results of the LightGBM and Random Forest classifications which can classify the Dhoho airport development land into the bare soil class well when compared to other models (see Figures 3-3). A tree-based model is robust because each decision tree will be trained with a different random subset of data. Then the voting results of all trees will be used as final predictions. The next stage is to validate the classification results of the best model (LightGBM). Validation of classification results has been carried out in areas that have many sugarcane plantations. Figure 3-4 shows the results of classification in a particular area. Areas 1, 2, and 3 are areas in Purwoasri sub-district that contain several sugarcane plantations, the difference between sugarcane plantations, open fields (rice fields, corn fields, etc.), trees and built-up area can be clearly captured by the LightGBM model.

3.3 Estimated Area of Sugar Cane Plantation

Data from official statistics on sugarcane plantation area is only available until 2021, Consequently, the estimated area of sugarcane plantations will be compared using the best model along with official statistic data sourced from the BPS and the Directorate General of Plantation. The comparison of the results of the estimated area of sugarcane plantations can be seen in Table 3-5 below.

Table 3-5: Comparison of the estimated area of sugarcane plantations.

	LightGBM	Official Statistic	Difference
Kediri Municipality	792.50	1,469	682.77
Kediri Regency	23,774	28,000	4,115
Total	24,671.23	29,469	4.797,77

According to the obtained results, the area of sugarcane plantations for the Kediri Regency and Kediri Municipality in 2021 is 23,885 Ha and 786.23 Ha, but the area of sugarcane plantations according to Official Statistic data is 28,000 Ha and 1,469 Ha. There is a total disparity of 4,797.77 hectares; this is likely attributable to the gathering of smallholder plantation data, which is susceptible to overestimate because it relies on estimates from informants and field officers (Ruslan & Prasetyo, 2021).

Furthermore, the estimated area of sugarcane plantations for September 2022 is also carried out. In September 2022, it is anticipated that the area of sugarcane plantations in the Kediri Regency and Kediri Municipality will be 18,897.6 ha and 571.87 ha, respectively. However, estimates from remote sensing tend to be underestimated when many areas are covered by clouds.

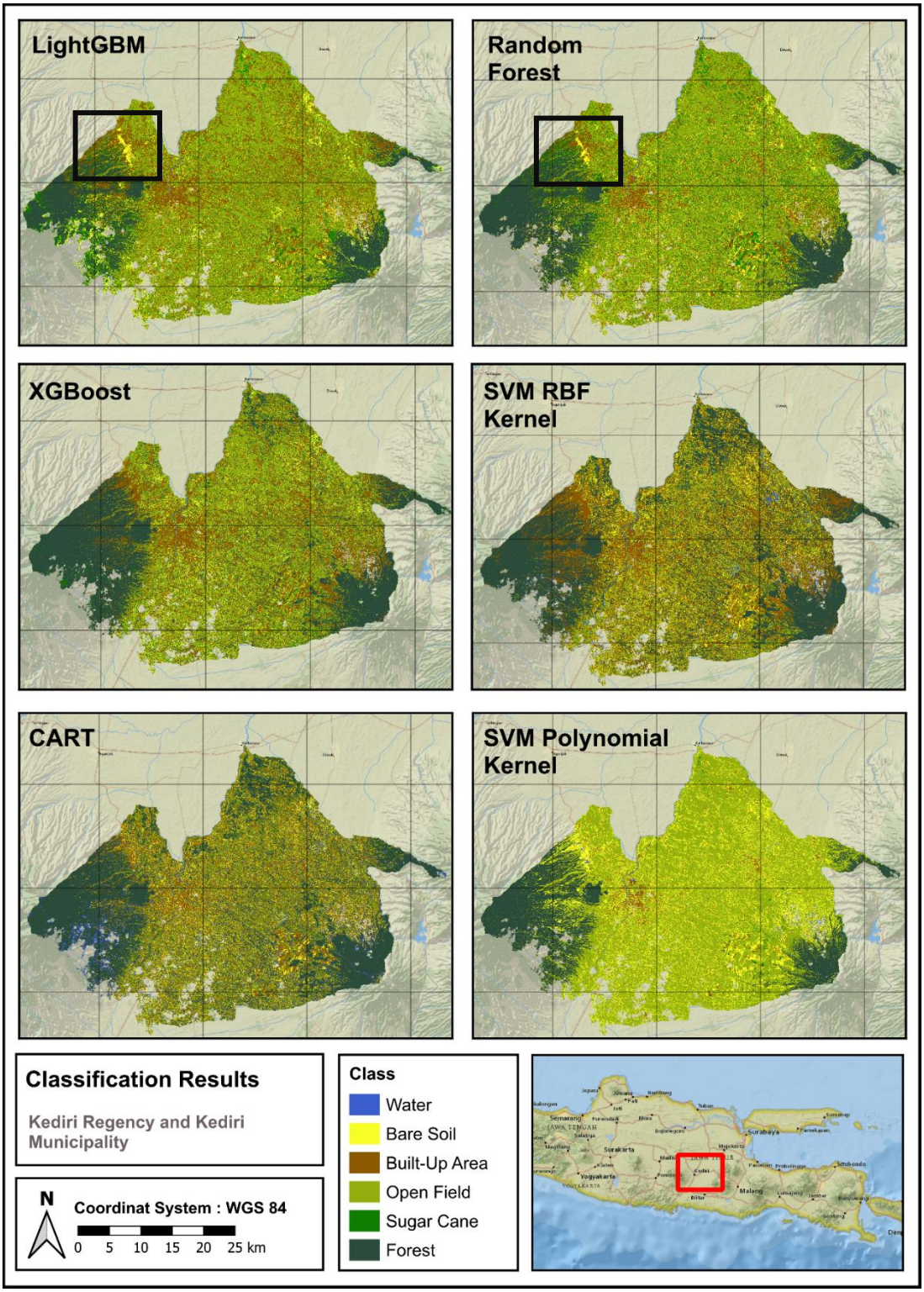


Figure 3-3: Classification Results for all six algorithms (LightGBM, Random Forest, XGBoost, SVM RBF Kernel, SVM Polynomial Kernel, and CART).

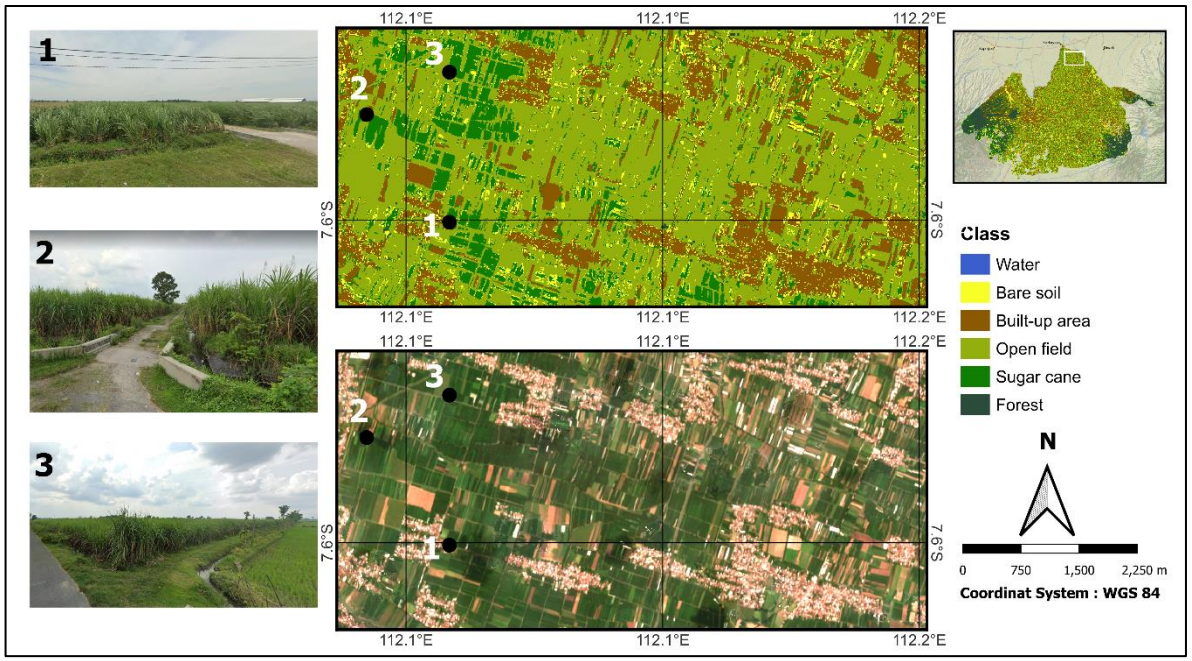


Figure 3-4: Classification result using LightGBM in specific area.

4 CONCLUSIONS

Sugarcane plantations in the Kediri Area can be identified well using the composite index values of NDVI, NDWI, NDBI, EVI, and elevation. The findings of hyper-parameter tuning with random search and stratified 10-fold cross validation indicate that LightGBM is the best model with accuracy and kappa values of 98.0% and 97.7%. Using the LightGBM model, the estimated area of sugarcane plantations in the Kediri Regency and Kediri Municipality in 2021 is 23,885 Ha and 786.23 Ha, there is a total difference of 4,797.77 ha; this is likely owing to the subjective nature of informants' and data collectors' estimates in the collection of smallholder plantation data, which makes it prone to overestimate. While 18,897.6 Ha and 571.87 Ha, respectively, are the findings of the estimation for the Kediri Area in September 2022.

ACKNOWLEDGEMENTS

We would like to thank the Polytechnic of Statistics for facilitating this research. Furthermore, we also appreciate the reviewers and editors of this paper for their constructive comments.

AUTHOR CONTRIBUTIONS

Authors: Ridson Al Farizal Pulungan (RAP) and Rani Nooraeni (RN). RAP & RN conceptualized the research objectives. RAP carried out the code writing, experimental simulations, analysis, visualization, and writing original draft preparation. RN selected the methodology, supervised the project, review and editing the manuscript. All authors have read and agreed to the published version of the manuscript.

REFERENCES

BPS (2022). *Statistik Tebu Indonesia 2021. Badan Pusat Statistik*, Indonesia.

Cevallos, J. C., Villagomez, J. A., & Andryshchenko, I. S. (2019). Convolutional neural network in the recognition of spatial images of sugarcane crops in the Tropical region of the coast of Ecuador. Procedia Computer Science, 150, 757-763.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd

acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.

Everingham, Y. L., Lowe, K. H., Donald, D. A., Coomans, D. H., & Markley, J. (2007). Advanced satellite imagery to classify sugarcane crop characteristics. Agronomy for sustainable development, 27, 111-117.

Ghaddar, B., & Naoum-Sawaya, J. (2018). High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research*, *265*(3), 9931004.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.

Indrawanto, C., Purwono, S., Syakir, M., & Rumini, W. (2010). *Budidaya dan pasca panen Tebu. ESKA media*. Jakarta.

Jiang, H., Li, D., Jing, W., Xu, J., Huang, J., Yang, J., & Chen, S. (2019). Early season mapping of sugarcane by applying machine learning algorithms to Sentinel-1A/2 time series data: a case study in Zhanjiang City, China. Remote Sensing, 11(7), 861.

Kementrian Pertanian. (2013). *Pedoman Pelaksanaan Pengelolaan Data Komoditas Perkebunan* (PDKP). *Direktorat Jenderal Perkebunan Kementerian Pertanian, Jakarta*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Kuhn, M., & Silge, J. (2022). Tidy Modeling with R. " O'Reilly Media, Inc.".

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. International Biometric Society, 33(1), 159–174.

Łoś, H., Mendes, G. S., Cordeiro, D., Grosso, N., Costa, H., Benevides, P., & Caetano, M. (2021, July). Evaluation of XGBoost and LGBM performance in tree species classification with sentinel-2 data. In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS (pp. 5803-5806). IEEE.

Luciano, A. D. S., Picoli, M. C. A., Rocha, J. V., Franco, H. C. J., Sanches, G. M., Leal, M. R. L. V., & Maire, G. L. (2018). Generalized space-time classifiers for monitoring sugarcane areas in Brazil. *Remote sensing of environment*, *215*, 438-451.

Marsuhandi, A. H., Triscowati, D. W., & Wijayanto, A. W. (2020). *Tinjauan Pemanfaatan Big Data Penginderaan Jauh Dan Pembelajaran Mesin Untuk Official Statistics di Wilayah Perkotaan. Jurnal Aplikasi Statistika & Komputasi Statistik*, 12(2), 31-40.

McCarty, D. A., Kim, H. W., & Lee, H. K. (2020). Evaluation of light gradient boosted machine learning technique in large scale land use and land cover classification. Environments, 7(10), 84.

Mulianga, B., Bégué, A., Clouvel, P., & Todoroff, P. (2015). Mapping cropping practices of a sugarcane-based cropping system in Kenya using remote sensing. *Remote Sensing*, *7*(11), 14428-14444.

Nurmasari, Y., & Wijayanto, A. W. (2021). Oil Palm Plantation Detection in Indonesia Using Sentinel-2 and Landsat-8 Optical Satellite Imagery (Case Study: Rokan Hulu Regency, Riau Province). International Journal of Remote Sensing and Earth Sciences (IJReSES), 18(1), 1-18.

Nonato, R. T., & Oliveira, S. R. D. M. (2013). Data mining techniques for identification of sugarcane crop areas in images of Landsat 5. *Engenharia Agricola*, *33*, 1268-1280.

Pramana, S., Yuniarto, B., Mariyah, S., Santoso, I., & Nooraeni, R. (2018).

Data mining dengan R konsep serta implementasi. Jakarta: InMedia.

Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold crossvalidation on ML classifiers for predicting cervical cancer. Frontiers in Nanotechnology, 4, 972421.

Ruslan, K., & Prasetyo, O. R. (2021). *Perbaikan Data Perkebunan Indonesia: Kopi, Gula dan Kakao*.

Schultz, B., Immitzer, M., Roberto Formaggio, A., Del'Arco Sanches, I., José Barreto Luiz, A., & Atzberger, C. (2015). Self-guided segmentation and classification of multi-temporal Landsat 8 images for crop type mapping in Southeastern Brazil. Remote Sensing, 7(11), 14482-14508.

Som-ard, J., Atzberger, C., IzquierdoVerdiguier, E., Vuolo, F., & Immitzer, M. (2021). Remote sensing applications in sugarcane cultivation: A review. *Remote sensing*, *13*(20), 404

Strobl, C., Boulesteix, A. L., & Augustin, T. (2007). Unbiased split selection for classification trees based on the Gini index. Computational Statistics & Data Analysis, 52(1), 483-501.

Sulaiman, A. A., Subagyono, K., Soetopo, D., Richana, N., Syukur, M., & Ardana, I. K. (2018). *Menjaring investasi meraih swasembada gula*. IAARD Press.

Syathori, A. D., & Verona, L. (2020). *Faktor-Faktor Yang Mempengaruhi Produksi Usaha tani Tanaman Tebu di Desa Majangtengah Kecamatan Dampit Kabupaten Malang. AGRIEKSTENSIA: Jurnal Penelitian Terapan Bidang Pertanian*, 19(2), 95-103.

Tariq, A., Yan, J., Gagnon, A. S., Riaz Khan, M., & Mumtaz, F. (2023). Mapping of cropland, cropping patterns and crop types by combining optical remote sensing images with decision tree classifier and random forest. Geo-Spatial Information Science, 26(3), 302-320.

Verma, A. K., Garg, P. K., & Hari Prasad, K. S. (2017). Sugarcane crop identification from LISS IV data using ISODATA, MLC, and indices-based decision tree approach. Arabian Journal of Geosciences, 10, 1-17.

Wang, M., Liu, Z., Baig, M. H. A., Wang, Y., Li, Y., & Chen, Y. (2019). Mapping sugarcane in complex landscapes by integrating multi-temporal Sentinel-2 images and machine learning algorithms. Land use policy, 88, 104190.

Wang, J., Xiao, X., Liu, L., Wu, X., Qin, Y., Steiner, J. L., & Dong, J. (2020). Mapping sugarcane plantation dynamics in Guangxi, China, by time series Sentinel-1, Sentinel-2 and Landsat images. *Remote Sensing of Environment*, *247*, 111951.

Zheng, H., Mahmoudzadeh, A., Amiri-Ramsheh, B., & Hemmati-Sarapardeh, A. (2023). Modeling Viscosity of CO2–N2 Gaseous Mixtures Using Robust Tree-Based Techniques: Extra Tree, Random Forest, GBoost, and LightGBM. ACS omega, 8(15), 13863-13875.