Classification of Paddy Growth Phase with Machine Learning Algorithms to Handle Imbalanced Multi-Class Big Data

H Suryono[1,2], H Kuswanto[1,*], N Iriawan[1]

[1]Department of Statistics, Faculty of Mathematics, Computation and Data Science Institut Teknologi Sepuluh Nopember,Surabaya 60111, East Java, Indonesia
[2]BPS-Statistics Indonesia, Jl. Dr. Sutomo 6-8, Jakarta 10710, Indonesia

*Corresponding author's e-mail: heri_k@statistika.its.ac.id

Abstract. The global Sustainable Development Goals (SDGs) adopted by countries in the world have significant implications for national development planning in Indonesia in the period 2015 to 2030. The Agricultural sector is one of the most important sectors in the world and has a very important contribution to achieving the goals. Availability of accurate paddy production data must be available to measure the level of food security. This can be done by monitoring the growth phase of paddy and predicting the classification of its growth phase accurately and precisely. The paddy growth phase has 6 classes with the number of class members usually not the same (imbalanced data). This study describes the results of the classification of paddy growth phases with imbalanced data in Bojonegoro Regency, East Java in 2019 using machine learning algorithms on the Google Earth Engine (GEE) platform. Classification is done by Classification and Regression Tree, Support Vector Machine, and Random Forest. Oversampling technique is used to deal the problem of imbalanced data. The Area Sampling Frame survey in 2019 conducted by BPS was used as a label for classification model training. The results showed that the overall accuracy (OA) using the Random Forest algorithm by modifying the dataset using oversampling was 82.30% and the kappa statistic was 0.76, outperforming the SVM and CART algorithms.

1. Introduction

The agricultural sector is one of the vital sectors in the world and Indonesia because it has a very significant contribution to the achievement of the goals of the Sustainable Development Goals (SDGs) and National programs. The implementation of national food security takes into account 3 (three) main components that must be met, namely: (1) Availability of sufficient and equitable food; (2) effective and efficient food affordability; and (3) Consumption of diverse and nutritionally balanced foods. The availability of accurate food data is very important to measure the level of food security. Monitoring of food crops through estimation of classification in the growth phase is carried out to answer the objectives of the SDGs.

Various fields and disciplines make use of remote sensing because of the ease of access and availability of data sets. In agriculture, remote sensing data is used to monitor paddy growth to ensure harvest [1], classification of plant species [2], estimation of harvest area [3], rice mapping [4], classifying rice plant phases [5]. BPS has monitored the paddy growth phase using the area sample frame survey (ASF) [6].

Landsat-8 satellite image data is one of the remote sensing data used for the Classification of Paddy Growth Phase. The segment approach in ASF is carried out on Landsat-8 data in the form of pixels

and produces a very large and unstructured amount of data which leads to the "Big Data" problem [7]. This problem demands new technologies and resources capable of handling large amounts of satellite imagery, such as cloud computing [8, 9]. In the data, it is proposed to use machine learning, because it will be difficult to find the model manually [10].

In the paddy growth phase which has 6 classes, namely: early vegetative, late vegetative, generative, harvesting, preparation, crop failure/puso, the number of class members is usually not the same [11], which means the class distribution is not uniform [12]. These data conditions cause machine learning algorithms to be biased towards imbalanced data problem, because the classifier will tend to predict the main class and ignore the small class. As a result, the prediction accuracy for the minority class will be much lower than for the majority class [13]. Given the significance of the class imbalance problem, this research aims to find the best model to solve the unbalanced data problem, which is the most common difficulty in machine learning. To overcome this problem, the data needs to be processed first to build a balanced dataset [14]. The approach taken is to apply the oversampling technique so that it has the potential to improve overall predictive performance. Therefore, we propose Machine learning methods, namely: Classification and Regression Tree (CART), Support Vector Machine (SVM) and Random Forest (RF) with oversampling using the Google Earth Engine (GEE) cloud computing platform to classify paddy growth phase from Landsat-8 satellite imagery.

2. Study Area

Geographically, Bojonegoro Regency is 112º 25’ until 112 º 09’ East longitude and 6 º 59’ until 7º 37’ South latitude. Bojonegoro Regency is part of East Java province on path/row 119/065 in Landsat-8 imagery, situated approximately 110 km to the East of Surabaya. Bojonegoro Regency area is a land area of 230.706 Ha. Bojonegoro Regency administration area consists of 28 districts and 430 villages. Land use in Bojonegoro Regency are paddy land 32.65%, dry land 24.39%, forest 42.74%, plantation 0.04%, others 0.18%. Across the eastern border of Bojonegoro is the Lamongan Regency, to the north is Tuban while to the south is Ngawi, Madiun, Nganjuk and Jombang. Blora is located to the west, in Central Java (Figure 1).

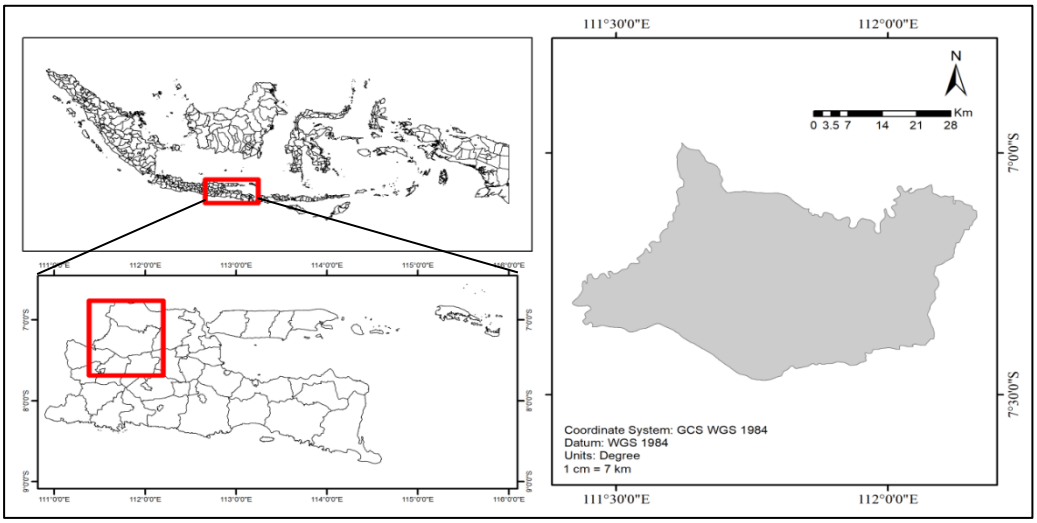

Figure 1. Geographical Location of Bojonegoro Regency

3. Reference Data

This study used 4 types of data: Landsat-8 data from GEE obtained in 2019, administrative boundary shapefiles, Area Sampling Frame (ASF) data from Statistics Indonesia (BPS), and land cover maps from Ministry of Environment and Forestry (KLHK). The ASF sample is used as a label for the training data based on the study area from the Landsat-8 satellite imagery map. The sample area in the ASF is called a segment with a size of 300 m x 300 m. One segment consists of 9 sub-segments measuring 100 m x 100 m.

3.1. Landsat archive data in the GEE

Landsat-8 is an Earth observation satellite from collaboration between NASA and the United States Geological Survey (USGS). This satellite consists of 9 Operational Land Image (OLI) sensors and 2 Thermal Infrared Sensors (TIRS). OLI has nine classes of bands that operate in the wavelength range of 0.433-2,300 m and provide images with a maximum resolution of 15 m [9]. Each spectral band in satellite imagery will produce different reflectance values between locations. GEE is a geospatial technology innovation that provides online access to Landsat-8 data [15]. This study used 7 bands consisting of aerosol, blue, green, red, NIR, SWIR1, and SWIR2 for the classification of paddy growth phases (Table 1).

Table 1. Types and uses of the Landsat-8 band (used in the study).

Band name	Landsat-8 Spectral range (μm)	Band Applications
Aerosol	0.43-0.45	Coastal and aerosol studies
Blue	0.45-0.51	Bathymetric mapping, distinguishing soil from vegetation, and deciduous from coniferous vegetation
Green	0.53-0.59	Emphasizes peak vegetation, which is useful for assessing plant vigor
Red	0.63-0.67	Discriminates vegetation slopes
NIR	0.85-0.88	Emphasizes biomass content and shorelines
SWIR1	1.57-1.65	Discriminates moisture content of soil and vegetation; penetrates thin clouds
SWIR2	2.11-2.29	Improved ability to track moisture content of soil and vegetation and thin cloud penetration

3.2. Vegetation Indices (VI)

Vegetation Indices (VI) is an optical measurement of the greenness of the vegetation canopy, the composite properties of leaf chlorophyll and the cover of the vegetation canopy. Several studies have been conducted on the vegetation indices from satellite image data to determine the growth phase of paddy plants [1]. Vegetation indices to detect the growth phase of paddy include the Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Normalized Difference Built-Up Index (NDBI), Normalized Difference Water Index (NDWI). ρ_{NIR}, ρ_{SWIR}, ρ_{RED} are two-way surface reflectance factors for each band [16]. The equations are:

$$NDVI = \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + \rho_{RED}}, \tag{1}$$

$$EVI = 2.5\frac{\rho_{NIR} - \rho_{RED}}{\left(1 + \rho_{NIR} + 6\rho_{RED} - 7.5\rho_{BLUE}\right)}, \tag{2}$$

$$NDBI = \frac{\rho_{SWIR} - \rho_{NIR}}{\rho_{SWIR} + \rho_{NIR}}, \tag{3}$$

$$NDWI = \frac{\rho_{NIR} - \rho_{SWIR1}}{\rho_{NIR} + \rho_{SWIR1}}. \tag{4}$$

3.3. Area Sampling Frame (ASF)

Area Sampling Frame (ASF) is a method developed by BPS and BPPT to calculate the area of paddy harvested each month. This method is formed by using satellite image maps and paddy fields maps. Sampling was carried out on an area measuring 300 m x 300 m which is called a segment. One segment (sample area) consists of 9 sub-segments measuring 100 m x 100 m. Paddy growth phase data generated from ASF in all sub-segments can estimate harvested area [6]. The paddy growth phases recorded in the ASF consist of 6 categories of paddy growth phases and can be seen in Table 2.

Table 2. Definition of Paddy growth phase in ASF survey.

No	Paddy growth phase	Definition	Days after planting
1	Early vegetative	The paddy growth phase starts from the beginning of growth until the maximum tillers	1-35 days
2	Late vegetative	The late vegetative phase begins when the tillers grow, starting from the appearance of the first tillers until the maximum number of tillers is reached	35-55 days
3	Generative	The growth phase starts from panicle out, ripening, until before harvest	55-105 days
4	Harvesting	The phase when the paddy is being harvested or has been harvested	
5	Preparation	The phase in which the paddy fields begin to be cultivated in preparation for paddy growth	
6	Puso (crop failure)	If there is an attack by plant-disturbing organisms or a disaster so that paddy production is less than 11% of normal	

4. Methods

This study used the Random Forest method with oversampling on the GEE data processing to provide a solution to the Geo Big Data remote sensing and data imbalanced problem for paddy growth phase classification. We processed and computed satellite image data using Google Earth Engine (GEE) which enables smooth and fast cloud and parallel processing on Google servers. Figure 2 shows a conceptual framework for paddy growth phase classification.

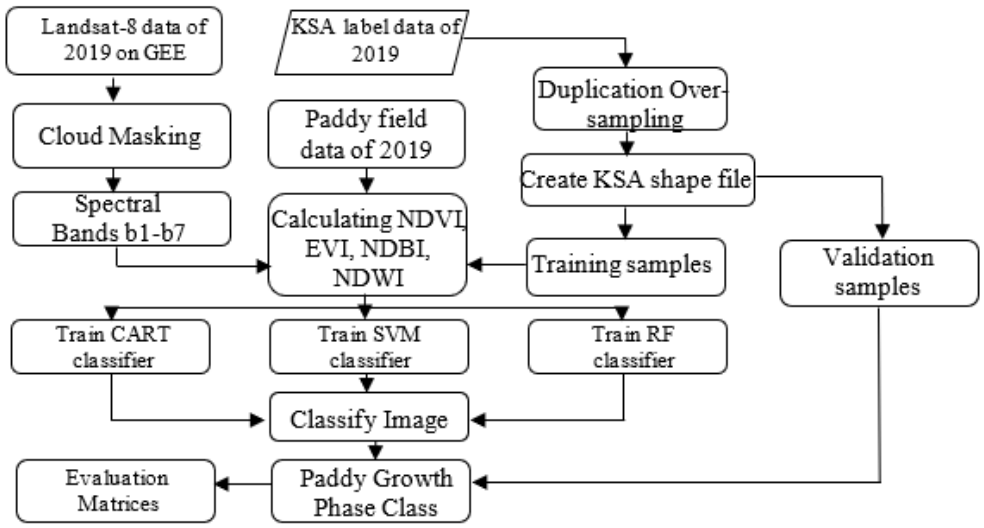

Figure 2. The conceptual framework

4.1. Google Earth Engine

Google Earth Engine (GEE) is a web portal that provides global time-series satellite imagery (over 40 years), cloud-based computing, and algorithms for processing data. Available data comes from several satellites: MODIS; U.S. Geological Survey's Landsat; National Oceanographic and Atmospheric Administration; European Space Agency's Sentinel. CART, RandomForest, NaiveBayes, and SVM are the only classifier packages available in Google Earth Engine for classification.

4.2. Duplication Oversampling

Oversampling is a sampling method that aims to balance the distribution of data by increasing the number of data in the minority class. The sampling method can be done by random or duplication. With the application of sampling on imbalanced data, the level of imbalance is getting smaller and classification can be done correctly [17]. In this paper, duplication oversampling is used by directly replicating all members of the minority class to approximate the number of members of the majority class.

4.3. Machine Learning Algorithms

Machine learning (ML) is a technique for inferring data with a mathematical approach. The essence of ML is to create mathematical models that reflect data patterns. The ML algorithm that will be used in this research is the Classification Regression Tree (CART), Support Vector Machine (SVM) and Random Forest (RF). Classification and Regression Tree (CART) is a nonparametric statistical method that can describe the relationship between the response variable (dependent variable) and one or more predictor variables (independent variable). This method can be used for classification trees using categorical type response variables and regression trees for continuous or numerical type response variables. Support Vector Machine (SVM) is a technique for making predictions, both in terms of classification and regression. SVM is used to find the optimal classifier function that can separate two data sets of two different classes. The ability of SVM has been studied by various researchers to classify data and show satisfactory performance [18]. Random forest (RF) is a grouping method based on decision tree ensemble (DT). RF is the development of the CART method by applying bagging and random feature selection to DT [19]. Classification decisions are made by majority vote among all trees.

4.4. Evaluation Matrices

The performance of prediction was evaluated as a function of accuracy, precision, and recall. The confusion matrix is a summary of the prediction result and performance measure for classification problems (Table 3). The number of true and false predictions for each class is summarized where the values (AP, AN) represent positive and negative test data, and the values (PP, PN) represent the predicted results for the positive and negative classes [20].

Table 3. Binary Confusion matrix.

	Actual Positive (AP)	Actual Negative (AN)
Predicted Positive (PP)	True Positives (TP)	False Positives (FP)
Predicted Negative (PN)	False Negatives (FN)	True Negatives (TN)

TP is data from the number of correct class member predictions in the positive class, TN is data from the number of correct class member predictions in the negative class, FP is data from the number of incorrect class member predictions in the positive class, FN is data from the number of incorrect class member predictions in the negative class.

4.4.1. *Accuracy.* Accuracy is the ratio of True (positive and negative) predictions to the overall data. Accuracy describes how accurate the model is in classifying correctly. Accuracy is an important measure used to assess the performance of a classification model. Accuracy is calculated as shown in Equation (5) as follows:

$$\text{accuracy:} \quad \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{5}$$

4.4.2. Precision. Precision is equal to the ratio of True Positive (TP) samples to the number of True Positive (TP) and False Positive (FP) samples. Precision is used to identify the amount of data that is correctly classified in the data set on the class. The precision is calculated as given in Equation (6) as follows:

$$\text{precision:} \quad \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{6}$$

4.4.3. Recall. Recall is the ratio of True Positive (TP) samples compared to the overall True Positive (TP) and False Negative (FN) sample data used to identify the number of correctly classified sample data in a class data set that can be predicted correctly. The Recall is calculated as given in Equation (7) as follows:

$$\text{recall:} \quad \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{7}$$

4.4.4. Kappa statistic. Kappa statistics are used to provide a quantitative measure of the magnitude of agreement between observers or the consistency between two measurement methods for nominal scales [21]. The kappa coefficient can measure the degree of agreement that classifies objects in mutually exclusive categories. The equation for kappa statistic (κ) is:

$$\kappa = \frac{\sum_{i=1}^{v} p_{ii} - \sum_{i=1}^{v} p_{i+} p_{+i}}{1 - \sum_{i=1}^{v} p_{i+} p_{+i}}, \tag{8}$$

where $\sum_{i=1}^{v} p_{ii}$ is the proportion of agreements and $\sum_{i=1}^{v} p_{i+} p_{+i}$ is the expected proportion of agreements by chance. v is the number of different nominal values for the performance indicator of interest.

5. Results

Bojonegoro Regency is one of the paddy barns in East Java Province. The ASF map and visualization of the results of the Classification of Paddy Growth Phases with the Random Forest Algorithm can be seen in Figure 3. Figure 4. is a map of the study area shaded in grey. Bojonegoro Regency is the study area in this research. Figure 5 shows the observation points in one sample segment (9 sub-segments). Each month the ASF survey counts as many as 110 segments. This means that the number of sub-segments every month is 992 sub-segments and a year is 11,912 sub-segments. A map of the results of the classification of the Paddy Growth Phase using the Random Forest Algorithm processed by the GEE platform is shown in Figure 6. Figure 3 and Figure 6 show the results of the classification using ROI and the area of paddy fields in Bojonegoro Regency, East Java. Thus, not all areas in Bojonegoro Regency as study areas will be classified.

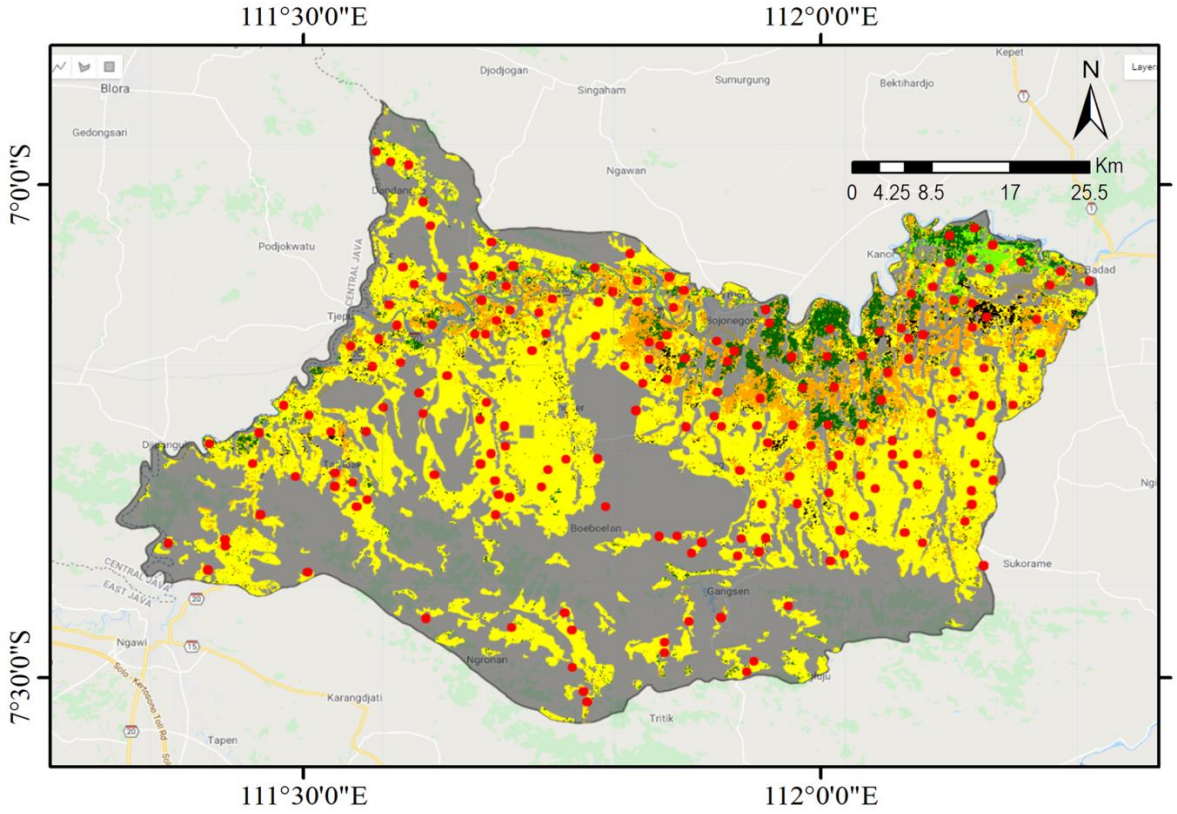

Figure 3. Map of ASF area and classification of Paddy Growth Phase with RF Algorithm.

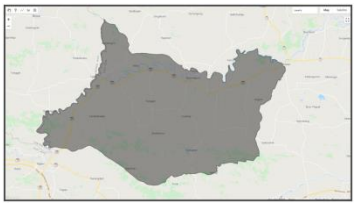

Figure 4. Bojonegoro Regency (study area)

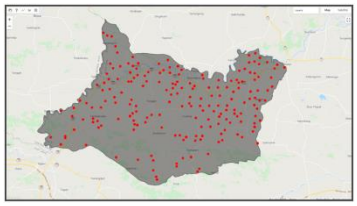

Figure 5. The distribution of ASF samples.

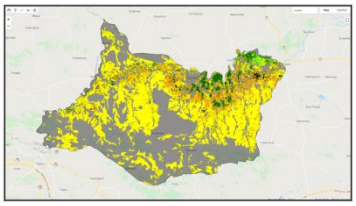

Figure 6. Classification results with RF algorithm using GEE (ROI: Paddy Area).

Table 4 shows the prediction performance for CART, SVM, RF with and without oversampling technique to handle imbalanced data. In the technique without oversampling, almost all machine learning algorithms show overall accuracy below 75%. CART shows a value of 60.95%, SVM shows a value of 70.79% and RF shows the highest result of 74.14%. The precision results for the three algorithms show that the lowest precision is shown by Class 5 (15.38% and 22.22%) on CART and SVM, while RF has higher precision for Class 5 although it is still in the 52% range. This shows that this model cannot predict Class 5 effectively because there are imbalanced data in the ASF data. Oversampling is applied to ASF data to handle imbalanced data so that the prediction performance of the minority class (Class 5) can be improved.

The oversampling technique used in this paper is duplication oversampling. For each algorithm, Table 4 presents the results in terms of accuracy, precision, and recall. The results show that the oversampling technique can improve accuracy and is predicted to be more efficient in terms of precision and recall for the minority class compared to the method without oversampling with an increase in precision in the CART algorithm by 64.62%, the SVM algorithm by 49.21% and the RF algorithm by 20.91. %. In the technique with oversampling, almost all machine learning algorithms

show an increase in overall accuracy compared to the technique without oversampling. CART showed an increase in accuracy to 71.78%, SVM accuracy increased to 78.57% and RF showed the highest accuracy increase, namely 82.30%. If you look at the precision results for the three algorithms, it can be seen that there is a very significant increase in all algorithms in Class 5, namely for CART, SVM, RF respectively by 80%, 71.43%, 76.47%%. This shows that the model has effectively handled the imbalance data to improve the prediction performance of the minority class.

Table 4. Performance of prediction for CART, SVM, RF with oversampling technique and without oversampling technique to handle imbalanced data

Algorithm	Class	Accuracy (%)	Precision (%)	Recall (%)
CART	0	60.95%	66.67%	76.92%
	1		50.00%	45.24%
	2		63.29%	49.02%
	3		67.14%	77.05%
	4		42.86%	50.00%
	5		15.38%	28.57%
SVM	0	70.79%	76.92%	90.91%
	1		66.67%	90.00%
	2		50.82%	68.89%
	3		91.86%	66.95%
	4		50.00%	60.00%
	5		22.22%	66.67%
RF	0	74.14%	80.95%	89.47%
	1		52.94%	77.14%
	2		66.67%	65.17%
	3		88.57%	77.50%
	4		53.85%	70.00%
	5		55.56%	62.50%
CART with oversampling	0	71.78%	52.17%	75.00%
	1		86.81%	71.17%
	2		50.63%	52.63%
	3		65.47%	79.82%
	4		66.67%	40.00%
	5		80.00%	51.61%
SVM with oversampling	0	78.57%	63.64%	100.00%
	1		70.00%	75.00%
	2		57.81%	66.07%
	3		82.28%	70.65%
	4		83.33%	83.33%
	5		71.43%	66.67%
RF with oversampling	0	**82.30%**	72.22%	86.67%
	1		86.81%	84.95%
	2		60.81%	75.00%
	3		92.14%	82.17%
	4		87.50%	82.35%
	5		76.47%	92.86%

The CART algorithm gives the highest improvement for the minority class classification, but overall the accuracy of the RF algorithm shows the best results in the classification, which is 82.30%.

It can also be seen in Table 4. that recall for class 5 (Puso) for the CART algorithm is only 28.57%. This shows that the CART model cannot effectively predict Class 5. While the recall for the SVM and RF algorithms has produced a Class 5 prediction above 60%, but it is still the lowest among the predictions of the other classes.

To clarify the performance of the proposed prediction, Table 5. summarizes the confusion matrix for each class in the paddy growth phase, namely early vegetative (0), late vegetative (1), generative (2), harvesting (3), preparation (4), failed harvest/puso (5) using RF algorithm with oversampling technique. The results show that our proposed method achieves the best performance in terms of prediction as a function of accuracy, precision, recall, and kappa statistics. The evaluation shows that our proposed method performs much better than other machine learning algorithms with or without oversampling, as shown in Table 4.

Table 5. Multi-class Confusion matrix of prediction of RF algorithm with oversampling technique.

		Actual Class						
		0	1	2	3	4	5	**Producer's Accuracy**
Predicted Class	0	13	3	2	2	3	0	72.22%
	1	0	82	4	0	0	0	86.81%
	2	0	10	48	12	1	2	60.81%
	3	0	0	11	133	2	1	92.14%
	4	0	0	0	0	15	0	87.50%
	5	0	0	0	2	0	19	76.47%
User's Accuracy		86.67%	84.95%	75.00%	82.17%	82.35%	92.86%	
						Overall Accuracy		**82.30%**
						Kappa Statistic		**0.76**

Table 5 shows that Class 5 can be mapped with good user's accuracy and producer's accuracy (92.86% and 76.47%, respectively). This shows that this model can classify paddy growth phases effectively with a kappa statistic of 0.76 which means model has a very good on the strength of agreement.

6. Conclusion

In summary, it can be concluded that oversampling is an appropriate technique to overcome the problem of data imbalance in the case of the paddy growth phase in Bojonegoro Regency, East Java. In this study, oversampling duplication technique is proposed to overcome the problem of data imbalance in classification using the ML algorithm. RF algorithm with oversampling has better performance than other ML classification algorithms without oversampling in terms of percentage accuracy for classifying paddy growth phases. This can be seen from the overall accuracy and kappa statistic (82.30% and 0.76). The model also succeeded in classifying the minority class (Class 5) contained in the ASF data efficiently. The performance of the proposed oversampling method is substantially better than without oversampling with imbalanced data on the 3 ML algorithm (CART, SVM, RF). Thus, our proposed method achieves the best performance in terms of paddy growth phase classification to support the contribution of achieving the SDGs goals in preparing valid food security data. Therefore, research on other models is an important next step in classifying Geo Big Data. Other methods, such as SVM and NN, can be used to improve model performance.

Acknowledgements

The first author would like to thank Statistics Indonesia (BPS) for providing the opportunity and support for the author to study in the Doctor Program. Furthermore, the authors also thank to other parties who contributed to the completion of this paper.

References

[1] Dirgahayu D and Made Parsa I 2019 Detection Phase Growth of Paddy Crop Using SAR Sentinel-1 Data *IOP Conf. Ser. Earth Environ. Sci.* 280

[2] Asgarian A, Soffianian A and Pourmanafi S 2016 Crop Type Mapping in a Highly Fragmented and Heterogeneous Agricultural landscape: A Case of Central Iran Using Multi-temporal Landsat 8 Imagery *Comput. Electron. Agric.* 127531–540

[3] You J, Li X, Low M, Lobell D and Ermon S 2017 Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data *31th AAAI Conf. Artificial Intelligence* 4559–4565

[4] Qiu B, Lu D, Tang Z, Chen C and Zou F 2017 Automatic and adaptive paddy rice mapping using Landsat images: Case study in Songnen Plain in Northeast China *Sci. Total Environ.* 598((11)):581–592

[5] Kim H O and Yeom J M 2014 Effect of red-edge and texture features for object-based paddy rice crop classification using RapidEye multi-spectral satellite image data *International Journal of Remote Sensing* 35 (19): 7046-7068

[6] Badan Pusat Statistik 2018 *Luas Panen dan Produksi Beras di Indonesia 2018* (Jakarta: BPS)

[7] Kussul N, Lemoine G, Gallego F J, Skakun S V, Lavreniuk M and Shelestov A Y 2016 Parcel-Based Crop Classification in Ukraine Using Landsat-8 Data and Sentinel-1A Data *IEEE J. Sel. Topics Appl. Earth Observ, Remote Sens.* vol 9 no 6 pp 2500–2508

[8] Shelestov A, Lavreniuk M, Kussul N, Novikov A and Skakun S 2017 Exploring google earth engine platform for big data processing: Classification of multi-temporal satellite imagery for crop mapping *Front. Earth Sci.* vol 5 p 17

[9] Mutanga O and Kumar L 2019 Google Earth Engine Applications *Remote Sens.* 11 591

[10] Dean J 2014 *Big Data, Data Mining and Machine Learning: Value Creation for Business Leaders and Practitioners* (New Jersey: John Wiley & Sons)

[11] Bak B A and Jensen J L 2016 High Dimensional Classifiers in The Imbalanced Case *Comp. Statistics & Data Analysis* 98 p 46-59

[12] Maurya C K, et al. 2016 Online sparse class imbalance learning on big data *Neurocomputing* 216 p 250-260

[13] Pouyanfar S and Chen S C 2017 Automatic Video Event Detection for Imbalance Data Using Enhanced Ensemble Deep Learning *International Journal of Semantic Computing* 11(1) p 85-109

[14] Belgiu M 2016 Random Forest in Remote Sensing: A Review of Applications and Future Directions *ISPRS Journal of Photogrammetry and Remote Sensing* 114 p 24-31

[15] Google Earth Engine 2021 *Google Earth Engine* Accessed: 9th August 2021 available from: https://earthengine.Google.com

[16] Huete A, Didan K, Miura T, Rodriguez E P, Gao X, Ferreira L G 2002 Overview of the radiometric and biophysical performance of the MODIS vegetation indices *Remote Sensing of Environment* 83 (1–2) 195–213.

[17] Laurikkala J 2001 Improving Identification of Difficult Small Classes by Balancing Class Distribution *Tech. Rep.* A-2001-2 University of Tampere

[18] Kuswanto H, Hidayati S, Salamah M, and Ulama B S 2017 Bootstrap resampling to detect active zone for extreme rainfall in Indonesia *The Asian Mathematical Conference 2016 (AMC 2016) Journal of Physics: Conf. Series* vol 893

[19] Breiman L 2001 Random forests *Machine learning* vol 45 no 1 pp 5–32

[20] Luque A, Carrasco A, Martín A and Heras A 2019 The impact of class imbalance in classification performance metrics based on the binary confusion matrix *Pattern Recognit.* 91 216–231

[21] Cohen J 1960 A coefficient of agreement for nominal scales *Educ Psychol. Meas.* 20:37-46