# VERIFYING PRACTICES OF REGENERATIVE AGRICULTURE: AFRICAN SMALLHOLDER FARMER DATASET FOR REMOTE SENSING AND MACHINE LEARNING

Yohei Nakayama[1], Grace Antwi[1], and Seiko Shirasaka[2]

[1]Degas Ltd.
[2]Keio University

## ABSTRACT

Despite Africa's contribution to global greenhouse gas (GHG) emissions being only a few %, the continent experiences the harshest impacts, particularly within its food production systems. Regenerative agriculture is receiving a large amount of attention as a method to strengthen both food security and climate change resilience in Africa. For practicing regenerative agriculture, carbon credits are issued, but verifying the methodologies on a large scale is one of the challenging points in popularizing it. In this paper, we provide a comprehensive dataset on regenerative agriculture in sub-Saharan Africa. The dataset has field polygon information and is labeled with several types of regenerative agriculture methodologies. The dataset can be applied to local site analysis, classification, and detection of regenerative agriculture with remote sensing and machine learning. We also highlight several machine learning models and summarize the baseline results on our dataset. We believe that by providing this dataset, we can contribute to the establishment of verification methods for regenerative agriculture. The dataset can be downloaded from https://osf.io/xgp9m/.

## 1 INTRODUCTION

Regenerative agriculture has received a large amount of attention in recent years as an alternative means of producing food that may have lower—or even net positive—environmental and/or social impacts (Newton et al., 2020). Carbon credits are issued for reducing GHG emissions through regenerative agriculture (Black et al., 2022). Thus, regenerative agriculture is considered a key to addressing climate change (Olsson et al., 2019). Methodologies of regenerative agriculture are mainly divided into five categories: (1) improve fertilizer (organic or inorganic) management, (2) improve water management/irrigation, (3) reduce tillage/improve residue management, (4) improve crop planting and harvesting (e.g., improved agroforestry, crop rotations, cover crops), (5) improve grazing practices (Black et al., 2022). In addition, adding biochar to soil in cropland is also treated as a methodology capable of issuing carbon credit (Etter et al., 2021). Monitoring and verifying the methodologies with various approaches such as surveys, use of sensors, on-site soil sampling, and remote sensing (Wang et al., 2022) are critical parts of the issuance of carbon credits. In particular, monitoring the practice implementations of regenerative agriculture or subsequent soil improvements through remote sensing and machine learning is receiving attention as it can perform periodic inspections in a scalable manner (Schreefel et al.; Ogungbuyi et al., 2023). However, open datasets related to regenerative agriculture are limited, and few provide polygons of each agricultural field. In this paper, we provide a dataset collected through a regenerative agriculture project in sub-Saharan Africa. The dataset includes field polygons of each agricultural field, labels of the type of tillage practiced, whether intercrop with cover crops is performed, and whether biochar is used. In total, 943 labeled samples are provided. We also applied several machine learning algorithms to the dataset and summarized the classification results.

## 2 Dataset Specification

The dataset was collected through a large-scale regenerative agriculture project in sub-Saharan Africa, particularly Ghana. As depicted in Figure 1, the study sites span the Northern and Upper West regions of Ghana. In total, the dataset includes field polygons of 938 agricultural fields. The latitude and longitude ranges are 11.0 to 9.0 and -0.65 to -2.6, respectively. On the sites, a combination of multiple types of farming methodologies, including several levels of reduced tillage, intercropping with maize and cover crops, and adding biochar into the soil, is practiced. Manual labels of each methodology and date of planting have been provided, which can be used for example for classification tasks. A summary of the category types and the number of samples for each category are shown in Table 1. The "None" value in cover crop types means intercropping is not conducted in the field, and only maize is planted. Regenerative agriculture has been conducted on the sites since 2023, and the date of planting ranges from June 2 to July 30.

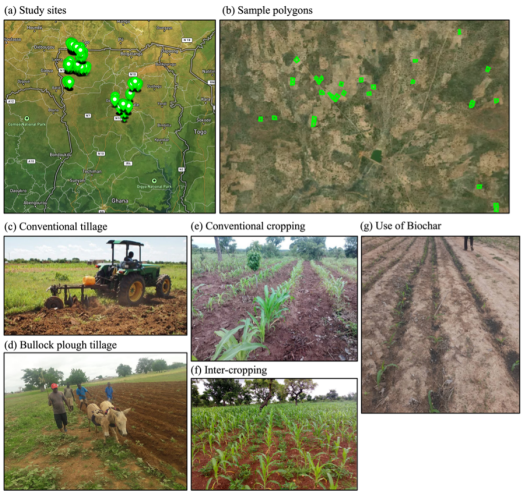


Figure 1: Schematic illustration of study sites in Ghana and field polygons included in the dataset. Sample pictures of each regenerative agriculture methodology.

Table 1: An overview of the presented dataset in this paper.

| Tillage methods | | Cover crop types | | Biochar usage | |
|---|---|---|---|---|---|
| Conventional | 449 | Bambara beans | 333 | True | 221 |
| No tillage | 235 | Pumpkin | 181 | False | 722 |
| Harrowed | 111 | Cowpea | 143 | | |
| Bullock plough | 77 | Soybeans | 71 | | |
| Ripping | 47 | Groundnuts | 50 | | |
| | | None | 160 | | |

## 3 Baselines and Experiments

### 3.1 Satellite Observation

The polygon information of the provided dataset can be used for local agricultural site analysis by mapping it with satellite observation data. In this paper, Sentinel-2 observations were used as the satellite data (Drusch et al., 2012). Observations of the target sites from March to September 2023 were collected, and only those with cloud coverage of 60% or less were extracted. For simplicity, we used the Normalized Difference Vegetation Index (NDVI) (Tucker, 1979) in this paper, which is usually referred to as a proxy for vegetation health monitoring (Bellón et al., 2017; Nay et al., 2018; Ayhan et al., 2020). In Sentinel-2, NDVI is calculated from the red and near-infrared bands. As the ground resolution of both bands is 10 m x 10 m, that of NDVI is also 10 m x 10 m.

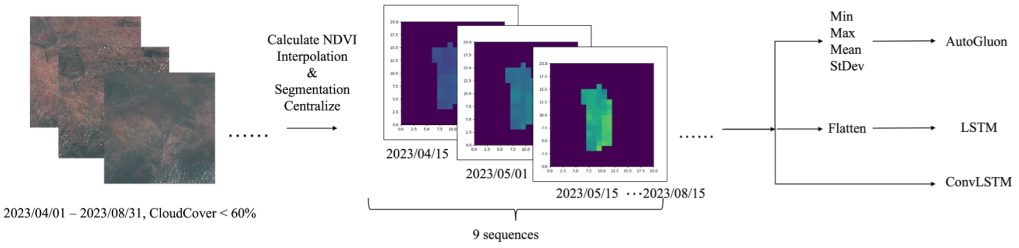


Figure 2: An overview of the collection process of satellite observation data, preprocessing, and model training.

### 3.2 Preprocessing

After calculating the NDVI values, we performed preprocessing before inputting them into machine learning models. Since the number of observation data collected and the observation date vary depending on the sites due to the cloud cover filter, we used interpolation to unify the number of observations and observation dates. The 1st and 15th dates of each month from April 15th to August 15th in 2023 are set as the reference dates, which is equivalent to plant and mid-season of Maize in Ghana. For a total of the 9 reference days, the NDVI value at each pixel was calculated by interpolating that from the previous and subsequent observation data. Lastly, we took each polygon of the dataset and segmented the pixels using the polygon, placing the polygon's centroid at the center of a 20 x 20 array of zeros. This array corresponds to an area of 200 m x 200 m from the ground resolution, and the area of all agricultural fields is smaller than that. Through this preprocessing, we created an input dataset where one data shape was 9 (sequence) x 1 (channel) x 20 (width) x 20 (height).

### 3.3 Baseline Models

We chose 3 algorithms to provide baselines of the dataset: AutoGluon, LSTM, and ConvLSTM, which are generally used for cropland mapping (Wang et al., 2022; Ahmad et al., 2023). With the 3 algorithms, we performed binary classification tasks. AutoGluon is an open-source AutoML framework that automates preprocessing, data splitting, model selection, and ensembling (Erickson et al., 2020). The base models of Autogluon (tabular task) are trained using bagging and multi-layer stack ensembling to improve classification or regression results. For AutoGluon, before feeding into the model, which calculates statistical values such as max, min, mean, and standard deviation of NDVI for each sample, so 9 (sequence) x 4 (statistical values of NDVI) = 36 columns were used as one input data sample. LSTM is one of the recurrent neural network (RNN) based architectures that is extensively used for classification or regression tasks with time-series satellite observation data (Tian et al., 2021; Sun et al., 2019). We also perform an additional process for LSTM, which flattens the input pixel, so 9 sequences of float values of 400 dimensions are used as input. The ConvLSTM is also a type of RNN for spatio-temporal prediction that has convolutional structures in both the input-to-state and state-to-state transitions (Shi et al., 2015). Unlike the original LSTM architecture, convolutional operations are used for internal matrix multiplications. Therefore, the module has the temporal modeling ability of an LSTM but can also capture local features, similar to a convolutional neural network (CNN). By applying ConvLSTM to the preprocessed data, it is possible to train a model by considering the distribution of NDVI values over farmland.

## 4 Result

For the 3 methodologies, we performed binary classification tasks with labels that are True/False of implementation of conventional tillage, intercropping with cover crops and biochar usage. Classifying whether each methodology is practiced on the field or not is the same as detecting the practice of that method. The results of our experiments are summarized in Table 2. We split the datasets into training/validation with a ratio of 8:2. In AutoGluon, we created a trained model by performing training with default settings and using the evaluation metric as balanced accuracy. AutoGluon tunes

Table 2: An overview of the presented dataset in this paper.

| **Tillage methods** | | | | |
|---|---|---|---|---|
| | Balanced Accuracy | F1 | Recall | Precision |
| AutoGluon | 0.60 | 0.64 | 0.72 | 0.58 |
| LSTM | 0.69 | 0.71 | 0.93 | 0.57 |
| ConvLSTM | 0.74 | 0.60 | 0.69 | 0.53 |
| **Intercropping** | | | | |
| | Balanced Accuracy | F1 | Recall | Precision |
| AutoGluon | 0.69 | 0.71 | 0.93 | 0.57 |
| LSTM | 0.54 | 0.48 | 0.64 | 0.54 |
| LSTM | 0.59 | 0.57 | 0.59 | 0.57 |
| **Biochar usage** | | | | |
| | Balanced Accuracy | F1 | Recall | Precision |
| AutoGluon | 0.74 | 0.60 | 0.69 | 0.53 |
| LSTM | 0.65 | 0.60 | 0.63 | 0.65 |
| ConvLSTM | 0.71 | 0.70 | 0.71 | 0.70 |

factors such as hyperparameters, early-stopping, and ensemble-weights in order to improve the evaluation metric on validation data. In LSTM and ConvLSTM, downsampling of the majority class is applied only to the training data to handle the class imbalance, and cross-entropy loss is used as the loss function. The model with the minimum loss value for the validation dataset during 50-epoch training is selected as the trained model. Balanced accuracy, F1 score, recall, and precision for the validation data are calculated using these trained models. For predicting tillage conditions, the maximum difference in F1 values between the models is 0.08, which is smaller than the differences in F1 values for other tasks. This suggests that the temporal evolution of NDVI is important in predicting tillage conditions and that spatial dispersion has little effect. AutoGluon's results are superior to others in predicting intercropping. This indicates that intercropping can be estimated just with temporal evaluation in NDVI and using statistical values. On the other hand, the two models that do not use statistical values but use raw data with high dimensions have lower accuracy. We conjecture that this is because the number of data samples is less than required to extract essential information for inference from the high dimensional input features. In predicting the use of biochar, the F1 score of ConvLSTM was about 0.1 higher than other results. This suggests that spatial dispersion of NDVI should be considered when predicting the use of biochar. Also, unlike the intercropping prediction, the fact that ConvLSTM also achieved high accuracy indicates that the number of data samples is sufficient to extract essential features for predicting the use of biochar from high-dimensional input.

## 5 Conclusion and Future Work

In this paper, we provided the first substantial dataset on regenerative agriculture and the baseline classification results that come with it. From the results of the experiment, we showed the effectiveness of the dataset for the detection and classification of the various methodologies. In the tillage condition and biochar usage, the dataset shows almost uniform results for all models. On the other hand, for intercropping prediction, although learning accuracy can be improved by calculating statistical values in advance and reducing the input dimension, it is suggested that the number of data samples is not sufficient when the input dimension is high. For future work, we plan to continually update the dataset from two perspectives. The first is to increase the sample size. The second is expanding the target sites. The footprint is limited to Ghana in this version, but it is planned to include the entire sub-Saharan region. Regarding the satellite observation data used, for simplicity, we only used NDVI calculated from Sentinel-2 observation. The classification accuracy of the tasks we proposed in this paper can be expected to improve by using more comprehensive multispectral observation data and high ground resolution data such as commercial observation data. In conjunction with the various types of satellite observations and this dataset, it's expected to develop state-of-the-art machine learning models that can contribute to the establishment of verification methods for regenerative agriculture.

## REFERENCES

Rehaan Ahmad, Brian Yang, Guillermo Ettlin, Andrés Berger, and Pablo Rodríguez-Bocca. A machine-learning based convlstm architecture for ndvi forecasting. *International Transactions in Operational Research*, 30(4):2025–2048, 2023.

Bulent Ayhan, Chiman Kwan, Bence Budavari, Liyun Kwan, Yan Lu, Daniel Perez, Jiang Li, Dimitrios Skarlatos, and Marinos Vlachos. Vegetation detection using deep learning and conventional methods. *Remote Sensing*, 12(15):2502, 2020.

Beatriz Bellón, Agnès Bégué, Danny Lo Seen, Claudio Aparecido De Almeida, and Margareth Simões. A remote sensing approach for regional-scale mapping of agricultural land-use systems based on ndvi time series. *Remote Sensing*, 9(6):600, 2017.

C Black, C Brummit, N Campbell, M DuBuisson, D Harburg, L Matosziuk, M Motew, G Pinjuv, and E Smith. Methodology for improved agricultural land management. *Available on: https://verra. org/metho dolog ies/vm0042-metho dology-for-impro vedagric ultur al-land-manag ement-v1-0/. Accessed*, 21, 2022.

Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

Hannes Etter, Andrea Vera, Chetan Aggarwal, Matt Delaney, and Simon Manley. Methodology for biochar utilization in soil and non-soil applications. *Verified Carbon Standard*, 2021.

John Nay, Emily Burchfield, and Jonathan Gilligan. A machine-learning approach to forecasting remotely sensed vegetation health. *International journal of remote sensing*, 39(6):1800–1816, 2018.

Peter Newton, Nicole Civita, Lee Frankel-Goldwater, Katharine Bartel, and Colleen Johns. What is regenerative agriculture? a review of scholar and practitioner definitions based on processes and outcomes. *Frontiers in Sustainable Food Systems*, 4:194, 2020.

Michael Gbenga Ogungbuyi, Juan P Guerschman, Andrew M Fischer, Richard Azu Crabbe, Caroline Mohammed, Peter Scarth, Phil Tickle, Jason Whitehead, and Matthew Tom Harrison. Enabling regenerative agriculture using remote sensing and machine learning. *Land*, 12(6):1142, 2023.

Lennart Olsson, Humberto Barbosa, Suruchi Bhadwal, Annete Cowie, Kenel Delusca, Dulce Flores-Renteria, Kathleen Hermans, Esteban Jobbagy, Werner Kurz, Diqiang Li, et al. Land degradation: Ipcc special report on climate change, desertification, land 5 degradation, sustainable land management, food security, and 6 greenhouse gas fluxes in terrestrial ecosystems. In *IPCC special report on climate change, desertification, land 5 degradation, sustainable land management, food security, and 6 greenhouse gas fluxes in terrestrial ecosystems*, page 1. Intergovernmental Panel on Climate Change (IPCC), 2019.

Loekie Schreefel, Rachel E Creamer, Hannah HE van Zanten, Evelien M de Olde, Annemiek Pas Schrijver, Imke de Boer, and Rogier PO Schulte. How to monitor the 'success' of (regenerative) agriculture: A perspective. *Available at SSRN 4525658*.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

Ziheng Sun, Liping Di, and Hui Fang. Using long short-term memory recurrent neural network in land cover classification on landsat and cropland data layer time series. *International journal of remote sensing*, 40(2):593–614, 2019.

Huiren Tian, Pengxin Wang, Kevin Tansey, Jingqi Zhang, Shuyu Zhang, and Hongmei Li. An lstm neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the guanzhong plain, pr china. *Agricultural and Forest Meteorology*, 310: 108629, 2021.

Compton J Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment*, 8(2):127–150, 1979.

Yu Wang, Han Liu, Lingling Sang, and Jun Wang. Characterizing forest cover and landscape pattern using multi-source remote sensing data with ensemble learning. *Remote Sensing*, 14(21):5470, 2022.