A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion

Thu Thuy Nguyen[a], Tien Dat Pham[b,d*], Chi Trung Nguyen[c], Jacob Delfos[d], Robert Archibald[d], Kinh Bac Dang[e], Ngoc Bich Hoang[f], Wenshan Guo[a], Huu Hao Ngo[a,f*]

[a]*Center for Technology in Water and Wastewater, School of Civil and Environmental Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia*
[b]*Department of Earth and Environmental Sciences, Macquarie University, North Ryde, NSW 2109, Australia*
[c]*Faculty of Science, Agriculture, Business and Law, UNE Business School, University of New England, Elm Avenue, Armidale NSW 2351, Australia*
[d]*Astron Environmental Services, 129 Royal Street, East Perth, Western Australia, 6004, Australia*
[e]*Faculty of Geography, VNU University of Science, 334 Nguyen Trai, Thanh Xuan, Hanoi, Viet Nam*
[f] *Institute of Environmental Sciences, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam*

***** Corresponding author: Huu Hao Ngo, *E–mail:* ngohuuhao121@gmail.com; Tien Dat Pham, Email: tiendat.pham@mq.edu.au

Abstract

Monitoring agricultural soil organic carbon (SOC) has played an essential role in sustainable agricultural management. Precise and robust prediction of SOC greatly contributes to carbon neutrality in the agricultural industry. To create more knowledge regarding the ability of remote sensing to monitor carbon soil, this research devises a state-of-the-art low cost machine learning model for quantifying agricultural soil carbon using active and ensemble-based decision tree learning combined with multi-sensor data fusion at a national and world scale. This work explores the use of Sentinel-1 (S1) C-band dual polarimetric synthetic aperture radar (SAR), Sentinel-2 (S2) multispectral data, and an innovative machine learning (ML) approach using an integration of active learning for land-use mapping and advanced Extreme Gradient Boosting (XGBoost) for robustness of the SOC estimates. The collected soil samples from a field survey in Western Australia were used for the model validation. The indicators including the coefficient of determination (R^2) and root - mean – square - error (RMSE) were applied to evaluate the model's performance. A numerous features computed from optical and SAR data fusion were employed to build and test the proposed model performance. The effectiveness of the proposed machine learning model was assessed by comparing with the two well-known algorithms such as Random Forests (RF) and Support Vector Machine (SVM) to predict agricultural SOC. Results suggest that a combination of S1 and S2 sensors could effectively estimate SOC in farming areas by using ML techniques. Satisfactory accuracy of the proposed XGBoost with optimal features was achieved the highest performance (R^2 = 0.870; RMSE= 1.818 tonC/ha) which outperformed RF and SVM. Thus, multi-sensor data fusion combined with the XGBoost lead to the best prediction results for agricultural SOC at 10 m spatial resolution. In short, this new approach could significantly contribute to various agricultural SOC retrieval studies globally.

Keywords: SOC, machine learning, multi-sensor data fusion, Sentinel 1, Sentinel 2

1. Introduction

Soil is one of the largest carbon pools in terrestrial ecosystems, and it plays a vital role in the global carbon cycles and care of the ecosystem (Lal, 2008; Zhou et al., 2020b). Agricultural soil organic carbon (SOC) contributes significantly to soil quality, soil fertility, agriculture and greenhouse gas emissions reduction by carbon sequestration in the agricultural SOC stock (Guo et al., 2021; Navarro-Pedreño et al., 2021; Venter et al., 2021). The agricultural SOC depends on land management practices, soil property and differs among rainfall zones (Guo et al., 2021; Six et al., 1998; Venter et al., 2021). Understanding the agricultural SOC distribution spatially is necessary to ensure food security and improve carbon sequestration in soil due to the increasing climate change problems (Gholizadeh et al., 2018). High-precision agricultural SOC data can help local authorities and governments establish appropriate strategies for agriculture and various farmland activities (Guo et al., 2021). Climate, ecological processes, agricultural production activities, soil characteristics, and land management are the key factors greatly influencing agricultural SOC.

The monitoring of agricultural SOC is complex due to the uncertainty of the above factors. Conventional SOC monitoring methods based on field experiments are time- and labour-consuming and subsequently, SOC mapping in large-scale areas is expensive (Forkuor et al., 2017). It is necessary to develop alternative approaches that are more cost-effective and accurate in predicting SOC. Numerous studies have attempted to solve this problem such as developing environmental models to improve the SOC estimation and applying remote sensing sensors to build digital SOC maps (Guo et al., 2021a; Guo et al., 2021b; Ha et al., 2021; He et al., 2021; Le et al., 2021; Mondal et al., 2017; Zhou et al., 2020). While developed SOC prediction models like a Full Carbon Accounting Model (FullCAM) or De-Nitrification De-Composition (DNDC) need a large amount of information from soil type, farming practices, and climate, they have illustrated their limitations in the prediction.

Recent advances in geospatial methods using earth observation (EO) datasets and advanced machine learning (ML) techniques can be effective in SOC monitoring (Vaudour et al., 2019). The use of multispectral, hyperspectral, or synthetic aperture radar (SAR) data from space-borne, air-borne remote sensing platforms, or unmanned aerial systems (UASs) has emerged as an innovative solution to address the issues of SOC prediction on farming lands. Although the performance of airborne RS and UAS with high spatial resolutions of hyperspectral images and extensive spectral information in SOC prediction outperforms the space-borne sensors with multispectral bands, the scarcity and high cost of hyperspectral data hinder their application in large-scale agricultural SOC estimation (Angelopoulou et al., 2019; Guo et al., 2021; see Table 1).

Table 1. Prediction performance of agricultural SOC in the recent literature.

Type of sensor	Sensor	ML Algorithm	R^2	Reference
Space-borne	Hyperion	PLSR	0.493	(Gomez et al., 2008)
	PRISMA	PLSR	0.51	(Castaldi et al., 2016)
	Landsat ETM+	ANN	0.63	(Mirzaee et al., 2016)
	S2	PLSR	0.56	(Vaudour et al., 2019)
	Gaofen 1	ELM	0.84	(Guo et al., 2020)
	S1+S2 +DEM	BRT	0.44	(Zhou et al., 2020b)
Air-borne	AHS160	SVM	0.89	(Stevens et al., 2010)
	HyMap	PLSR	0.85	(Vohland et al., 2017)
Unmanned Aerial Systems	Mini-MCA6	SVM	0.95	(Aldana-Jague et al., 2016)

PLSR: Partial Least Squares Regression; SVM: Support Vector Machines; ANN: Artificial Neural Networks; ELM: Extreme Learning Machine; BRT: Boosted Regression Trees;

Multispectral remote sensing sensors such as Hyperion, S-2, S-1, Gaofen 1, Landsat ETM+, and PRISMA have demonstrated their usefulness in agricultural SOC estimation. The free-of-charge multispectral images are an effective solution to address the problems concerning hyperspectral images in agricultural SOC monitoring. Gaofen 1 - launched by China National Space Administration – has great potential in estimating agricultural SOC with 0.84 R^2 compared to other multispectral images (Guo et al., 2020). However, its spectral bands are not widely supported by various agencies of the Chinese government. Combining multi-sensors in predicting agricultural SOC has been done in recent studies such as: the integration of Sentinel 1 and Sentinel 2; and joining Sentinel 2 and Sentinel 3 (Zhou et al., 2020b; Zhou et al., 2021). Multi-sensor data fusion technology is a promising way to improve prediction performance compared to single sensor technology (Khaleghi et al., 2013; Le et al., 2021).

A few studies have combined optical data (S-2) and SAR data (S-1) to estimate agricultural SOC content (Zhou et al., 2020) . Recently, Zhou et al (2020) explored the potential of using S1, S2, and digital elevation model (DEM) data in predicting agricultural SOC by Boosted Regression Tree (BRT) machine learning technique. It had a prediction accuracy of 0.44 R^2, which is quite low compared to other research (Table 1). It is likely due to the optimisation of hyper-parameters tunning and the selection of predictor variables during the construction phase of the ML techniques. A range of ML algorithms were used for agricultural SOC monitoring which are presented in table 1. The XGBoost was used in many studies due to its high predictive performance and being an effective supervised learning algorithm for addressing various classification and regression tasks with promising results (Chen and Guestrin, 2016), however; it has not been applied for agricultural SOC monitoring. For these reasons, the present study aims to develop a novel framework using free-of-charge multi-sensor Sentinel 2 and Sentinel 1 with state-of-the-art extreme gradient boosting (XGBoost) to predict agricultural SOC stocks. The specific objectives are to: (1) assess the feasibility of using multi-spectral images and SAR dataset in estimating agricultural SOC; (2) compare the prediction performance of the XGBoost to two other well-known ML techniques (random forest (RF) and support vector machine (SVM)) with various scenarios of data-fusion level in agricultural SOC prediction; and (3) highlight important predictor features in mapping agricultural SOC stock at 10 m spatial resolution. The novel agricultural SOC prediction framework will then be expanded so that relevant stakeholders are aware of the many advantages for agricultural management, climate change mitigation and landholders wanting to make more profit via carbon markets.

2. Materials and methods

2.1. Study area

The study sites are the Wests area which belongs to Goomalling shire (latitude coordinate: -31°18'S and longitude coordinate: 116° 49' E), and Cookies area which belongs to Northam shire (latitude: -31° 39' S, and longitude: 116° 39' E). These areas are located in the agricultural lands of Western Australia (WA). The agricultural sector plays an essential role in the WA’s economy. Pastoral and cropping are two main agricultural activities in the WA. According to Australian Bureau of Agricultural and Resource Economics, there are three key agricultural climatic zones in Australian, which are High-rainfall, Wheat-sheep, and Pastoral zones (Salim & Islam, 2010). While 95 per cent of gross value of agricultural production in the WA comes from the high-rainfall and wheat-sheep zones, only 5% of agricultural products is produced from pastoral zones. As the agricultural of the WA bases totally on rainfall, the main season for crop production in the WA is from April to October. The rainfall in growing season ranges between 146 to 294 mm (Petersen & Hoyle, 2016).

2.2. Soil samples collection

From very high spatial resolution Google Earth imagery and Sentinel 2 imagery, a total of 266 digitizing points for both vegetation and bare soil locations were selected to generate land-use binary maps (Figure 1). An Advanced ML technique with five-fold cross validation (CV) method were applied for binary land-use classification mapping. The classification accuracy of the XGBoost model were compared with the two well-known ML algorithm such as the RF and SVM technique. The overall accuracy, kappa coefficient, precision, recall and F1_score served as evaluation metrics. The best model with the highest value of overall accuracy, F1 score and Kappa coefficient was chosen to produce the binary land-use map. The binary land use classification map devised in the study areas served to identify bare-soil points for agricultural SOC sampling. The active learning technique in remote sensing classification was employed to assist in designing and sampling soil carbon, which helps minimise effects of vegetation on SOC contents (Fu et al., 2010; Tuia et al., 2011).

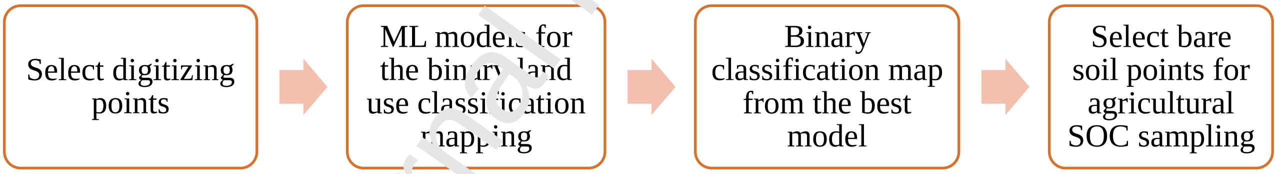
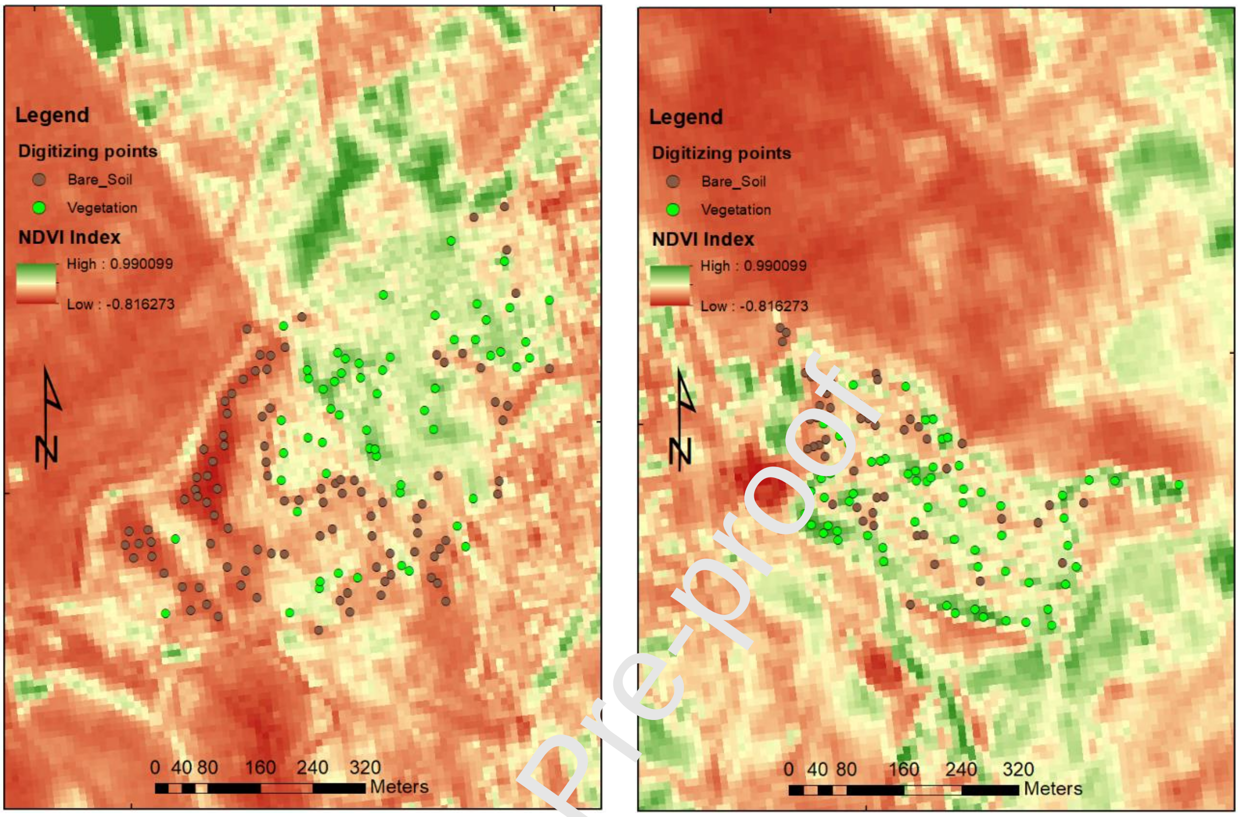

Select digitizing points
ML models for the binary land use classification mapping
Binary classification map from the best model
Select bare soil points for agricultural SOC sampling

Figure 1. Flow chart of land-use binary mapping and SOC samples selection using an active learning method

The agricultural SOC field survey was carried out in April 2021. Forty bare-soil sampling locations with a pixel (size of 10m x 10m) across the study areas (20 points for each area) were selected based on the binary map (Figure 2). A Differential Global Positioning System (DGPS) - a refined version of the Global Positioning System (GPS) - was used to identify precisely the samples' location with an accuracy of 1-3 cm (Michalski and Czajewski; 2004). Four soil cores were taken in each sampling plot. The dimensions of the core was 7 cm in depth and 7.3 cm in diameter. The total agricultural SOC of soil samples was analysed in the laboratory by Rayment and Lyons Method 6B1 (Heanes, 1984).



(a) Wests **(b) Cookies**

Figure 2. Study areas and digitizing point selection: (a) Wests, and (b) Cookies

2.3. Research framework

The research process includes four main phases (Fig. 3): (1) collection of surface soil dataset (0-10cm) based on the binary land-use map; (2) computation of predictor variables from optical (Sentinel 2) and synthetic aperture radar (Sentinel 1) remote sensing data; (3) spatial modelling of agricultural SOC based on advanced machine learning techniques including XGBoost, RF and SVM model; and (4) evaluating the model's performance with 70% of SOC dataset generated for models' training and 30% for models' testing. This was done to select the most accurate model for SOC prediction and mapping the spatial patterns of agricultural SOC.

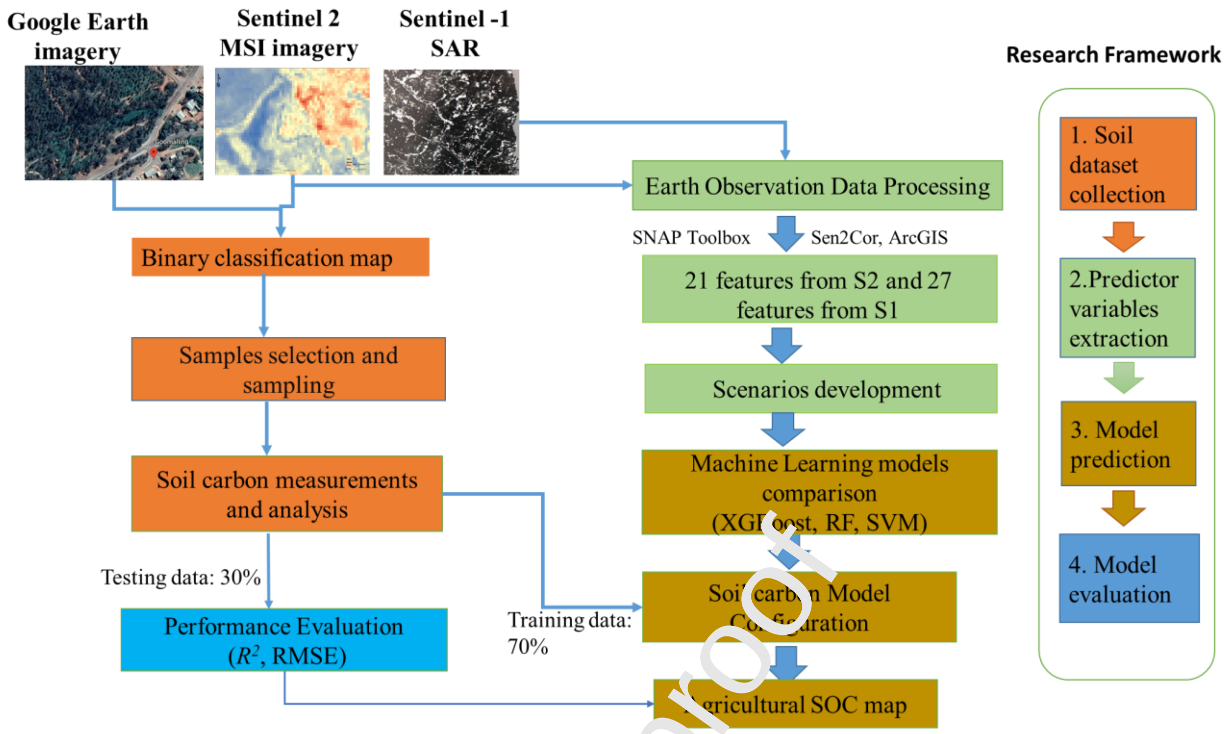

Google Earth imagery
Sentinel 2 MSI imagery
Sentinel -1 SAR
Research Framework
Earth Observation Data Processing
SNAP Toolbox
Sen2Cor, ArcGIS
Binary classification map
21 features from S2 and 27 features from S1
Samples selection and sampling
Scenarios development
Soil carbon measurements and analysis
Machine Learning models comparison (XGBoost, RF, SVM)
Testing data: 30%
Soil carbon Model Configuration
Training data: 70%
Performance Evaluation (R^2, RMSE)
Agricultural SOC map
1. Soil dataset collection
2.Predictor variables extraction
3. Model prediction
4. Model evaluation

Figure 3. A novel established framework of agricultural SOC prediction using multi-sensor data fusion

2.4. Remote sensing data acquisition and image processing

2.4.1. Data acquisition

In this study, S-2 multispectral satellite and S-1 C-band dual polarimetric SAR sensors computed the predictor indicators for agricultural SOC. Table 2 illustrates satellite data acquisitions from S-2A Multispectral Instrument (MSI) and S-1C Ground Range Detected (GRD) product with a dual-polarization data (Vertical transmit Vertical receiving (VV) and Vertical transmit Horizontal receiving (VH)). The ten multispectral bands from S2 sensor were employed with spatial resolution ranging from 10 to 20. Sentinel 1 and Sentinel 2 images were obtained from the Copernicus Open Access Hub from European Space Agency (ESA). The SNAP Sentinel Application Platform toolbox were employed for both optical and SAR data processing, whereas ArcGIS 10.3 was used to generate spatial agricultural SOC. The acquisition dates were closer to the field data collection dates (from 23 to 28 April 2021).

Sensor	Scene / Tile ID	Acquisition date (month/day/year)	Processing level	Spatial resolution (m)	Spectral band/ polarization
S-2	50JML	04/17/2021	1C	10 – 20	13 multispectral bands
S-1	S1B_IW_GRDH1SDV	04/27/2021	GRD	10	Dual- polarization (VV and VH)

Table 2. Satellite data acquisition for the study sites

Source: European Space Agency ESA, 2021

2.4.2. Image transformation of Sentinel 2 imagery

The level 1-C Sentinel 2 served as the top of atmosphere (TOA) processing level and were geocoded in the projection of World Geodetic System (WGS84) - Universal Transverse Mercator (UTM) zone 50 South (50S). Then the S-2 data was transformed to surface reflectance by the bottom of atmospheric (BOA) correction using the ESA Sen2Cor plugin in the SNAP (Louis at al., 2016). In this study, a total of ten relevant S-2 bands including B2, B3, B4, B5, B6, B7, B8, B8A, B11, and B12 were used for this study from thirteen original S-2 bands. The 10 bands of Sentinel 2 are employed extensively to evaluate soil properties (Elhag & Bahrawi, 2017). The 10 bands were resampled to a ground sampling distance (GSD) of 10 m. Vegetation and soil indices are mentioned as being sensitive to soil organic carbon content which recently were applied for soil attribute prediction (Jin et al., 2017). While seven vegetation indices (VIs) were computed by vegetation radiometric indices algorithms, four soil indices (SIs) were extracted from a soil radiometric indices function, which are derived from a Thematic Land Processing module in SNAP (Pasqualotto et al.,

2019) (Table 3). A total of 21 predictor variables derived from S-2 were used for agricultural SOC.

Table 3. Vegetation and soil predictor variables derived from Sentinel 2 (adapted from Pham et al.,,2020))

Vegetation and Soil Index	Acronyms	S-2 band wavelengths	References
Ratio Vegetation Index	RVI	$\frac{\text{NIR}}{\text{Red}}$	(Tucker 1979)
Normalized Difference Vegetation Index	NDVI	$\frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}$	(Rouse Jr et al. 1974)
Green Normalized Difference Vegetation Index	GNDVI	$\frac{\text{NIR} - \text{Green}}{\text{NIR} + \text{Green}}$	(Gitelson et al. 1996)
Normalized Difference Index using Bands 4 & 5 of S-2	NDI45	$\frac{\text{RE1} - \text{Red}}{\text{RE1} + \text{Red}}$	(Delegido et al. 2011)
Soil Adjusted Vegetation Index	SAVI	$(1 + L)(\frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red} + \text{L}})$ L = 0.5 in most conditions	(Huete 1988)
Inverted Red-Edge Chlorophyll Index	IRECl	$\frac{\text{RE3} - \text{Red}}{\text{RE1/RE2}}$	(Frampton et al. 2013)
Modified Chlorophyll Absorption in Reflectance Index	MCARI	$[(\text{RE1} - \text{Red}) - 0.2 \times (\text{RE1} - \text{Green})] \times (\text{RE1} - \text{NIR})$	(Daughtry et al. 2000)
Brightness index	BI	$\frac{\sqrt{(\text{Red} \times \text{Red}) + (\text{Green} \times \text{Green})}}{2}$	(Escadafal 1989)
Brightness index 2	BI2	$\frac{\sqrt{(\text{Red} \times \text{Red}) + (\text{Green} \times \text{Green}) + (\text{NIR x NIR})}}{2}$	(Escadafal 1989)

Vegetation and Soil Index	Acronyms	S-2 band wavelengths	References
Redness index	RI	$\frac{\text{Red} \times \text{Red}}{\text{Green} \times \text{Green} \times \text{Green}}$	(Mathieu et al. 1998)
Colour index	CI	$\frac{\text{Red} - \text{Green}}{\text{Red} + \text{Green}}$	(Mathieu et al. 1998)

Note: Band wavelengths of S-2: B2: Blue (492 nm), B3: Green (560 nm), B4: Red (665 nm), B5: Red-edge 1 (RE1) (704 nm), B6: Red-edge 2 (RE2) (740 nm), B7: Red-edge 3 (RE3) (783nm), B8: near-infrared (NIR) (833 nm), B8A: Narrow-NIR (865 nm), B11: short-wavelength infrared (SWIR1) (1614 nm), and B12: SWIR2 (2202 nm).

2.4.3. Image transformation of Sentinel-1 imagery

The extraction of Sentinel 1 data included eight steps which were conducted in the SNAP application using the Radar toolset to convert the S-1 C-band SAR raw intensity signal data to scale backscatter coefficient (σ^0) in decibel (dB) as suggested by Pham et al.,(2020) and Filipponi (2019). The steps includes: (1) Correct the orbit file; (2) Thermal and border noise removal; (3) Radiometric calibration; (4) Speckle filtering; (5) Range Doppler terrain correction; (7) Normalized radar backscattering coefficient by the equation 1 below; (8) S-1 SAR band transformation to create five predictor features including VV/VH; VH/VV; VV-VH; VH-VV; (VV+VH)/2; and (9) computation of 20 features using grey level co-occurrence matrix (GLMC) from S-1 VV and VH Polarizations (Fig 4).

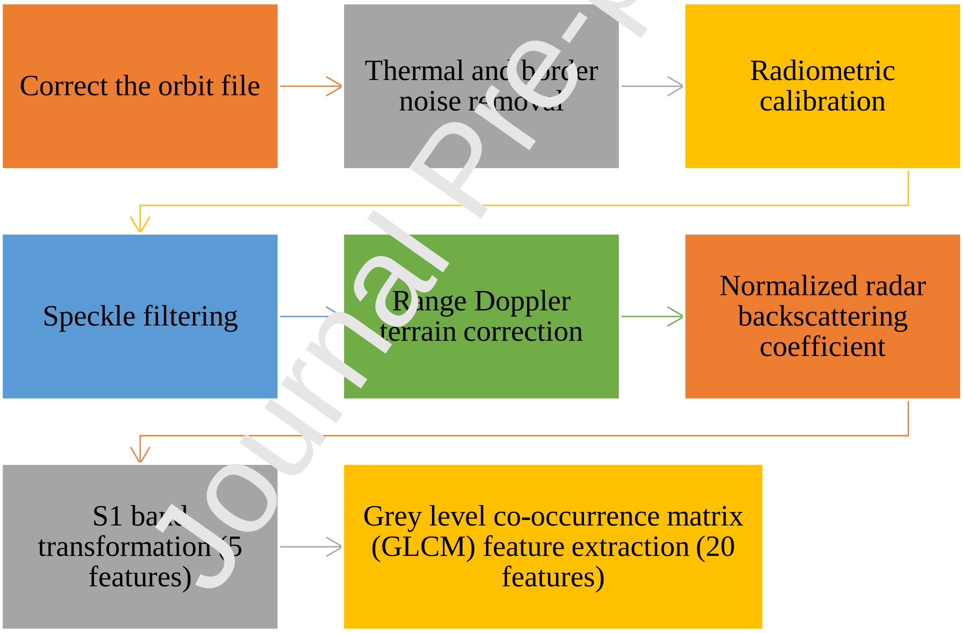

Correct the orbit file
Thermal and border noise removal
Radiometric calibration
Speckle filtering
Range Doppler terrain correction
Normalized radar backscattering coefficient
S1 band transformation (5 features)
Grey level co-occurrence matrix (GLCM) feature extraction (20 features)

Figure 4. Steps of Sentinel 1 pre-processing and processing

A total of 27 features were extracted and computed from Sentinel 1. These features contained: the two bands from dual polarization (VH and VV); the five SAR transformed bands (VV/VH; VH/VV; VV-VH; VH-VV; (VV+VH)/2); and the 20 new features extracted from VV and VH using the GLMC algorithm (VV_Contrast, VV_Dissimilarity, VV_Homogeneity, VV_Angular Second Moment, VV_Energy, VV_Maximum Probability, VV_Entropy, VV_GLCM Mean, VV_GLCM Variance, VV_GLCM Correlation, VH_Contrast, VH_Dissimilarity, VH_Homogeneity, VH_Angular Second Moment, VH_Energy, VH_Maximum Probability, VH_Entropy, VH_GLCM Mean, VH_GLCM Variance, and VH_GLCM Correlation).

2.4.4. Scenarios development

Scenarios were constructed based on the different number of predictor features and the combinations of sensors. While Scenario 1 and Scenario 2 were developed from S-2 derived predictors, Scenario 3 and Scenario 4 were built from S-1 derived predictors. Scenario 1 (SC1) included only 10 features from 10 S-2 bands. Scenario 2 (SC2) consisted of a total of 21 S-2 derived predictors including 10 S-2 bands, 7 VIs bands, and 4 SIs bands. Scenario 3 (SC3) and Scenario 4 (SC4) comprised 7 and 27 predictor features from the S-1 sensor, respectively. Scenario 5 (SC5) included all features based on the combination of S-2 and S-1. The purpose of scenarios development was to assess the impact of the type of predictor variables and the level of different features combinations on how well agricultural SOC prediction went.

2.5. Machine learning techniques

2.5.1. Extreme gradient boosting (XGBoost)

The XGBoost technique was introduced by Chen and Guestrin (2016). It shares the same theory with other gradient tree boosting algorithms. The XGBoost algorithm is described as a scalable end-to-end tree boosting which is a highly accurate machine learning technique and has widely applied to solve data mining problems (Chen and Guestrin, 2016). The novelty of XGBoost is its scalability in all scenarios so it can handle sparse data challenges. This advanced ML techniques is able to handle both classification and regression tasks (Ha et al., 2021b). The further merits of the XGBoost are parallelization, out-of-core computation, and cache optimization, which help the training process of the system more quickly than existing gradient boosted regression tree methods. This technique can easily deal with the problem of a model's complexity especially if it has a large dataset. Moreover, the XGBoost method can use integrated optimization algorithms to tune important hyper-parameters such as the number of trees and the rate of learning to suit a specific dataset. In this study, the best structure with 100 trees, and a learning rate set at 0.5 and gamma value of 5 was found the highest performance in the XGBoost model.

2.5.2. Random forest (RF)

The RF algorithm is one of the most popular machine learning algorithms, and it can be used effectively for a wide range of applications (Breiman, 2001; Pham et al., 2020). This technique includes a large number of regression trees. Each regression tree is built by the unique bootstrap sample from the original dataset, which decreases the sensitivity of the RF method to overfitting problems. Normally, the dataset will be divided with about two-thirds of the samples (in-bag data) for the training sets and the remaining samples for the test sets (Out-Of-Bag (OBB data). Two essential parameters including the number of regression trees and number of predictor variables must be defined in the RF model. In the current work, the RF model with 100 trees and the maximum number of 11 features had the highest performance for this study area.

2.5.3. Support vector machine (SVM)

Developed by Cortes and Vapnik (1995), the SVM algorithm is a well-known supervised learning technique based on the kernel approach and statistical theory, which can applied for classification, regression and outliers detection (Cortes & Vapnik, 1995; Cristianini & Ricci, 2008). While the SVM can help solve non-linear dataset, this method is not effective with a noisy and overlapped dataset. One of the advantages of SVM is that it can work accurately with a small number of training datasets. The SVM algorithm's performance is based on the selection of kernel functions and their parameters. There are three hyper-parameters in the SVM method including regularization parameter, the kernel function, and gamma controlling the overfitting. The hyper-parameters of the SVM method are fewer than other machine learning algorithms. Four kernel function types include polynomial, sigmoid, linear and radial basis function. In this study, the grid search with a five-fold CV was used to determine the optimal hyper-parameters of each ML algorithm in the Python environment. In this work, the SVM algorithm with the radial basis function (RBF) kernel and the *C* value of 10000 was used, and the epsilon value of 0.01 as the best values for tuning hyper-parameters of the SVM model.

2.6. Model performance evaluation

To assess the model performance of binary land-use classification, five evaluation criteria have been used including overall accuracy (*OA*), kappa coefficient (*KC*), precision (*P*), Recall (*R*), and F1 score (*F1*) (Chicco & Jurman, 2020; Ha et al., 2021).

For agricultural SOC retrieval, two common validation criteria were employed to assess the performance of machine learning techniques with different scenarios including: the root mean square error (RMSE), and the coefficient of determination (R^2). Superior model performance illustrates the higher R^2 and lower RMSE. These criteria are evaluated using the equations below:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \quad (1)$$

$$R^2 = \frac{\sum_{i=1}^{n}(P_i - \overline{O_i})}{\sum_{i=1}^{n}(O_i - \overline{O_i})} \quad (2)$$

Where: n indicates the number of soil samples; P_i and O_i illustrate the predicted SOC value and measured SOC value of the i sample, respectively.

3. Results

3.1. Land-use binary mapping

Land-use classification results found by the XGBoost,, the RF and the SVM algorithms are indicated in Table 4 below. The results present the high accuracy of land-use binary mapping at study sites using the S-2 dataset. The XGBoost algorithm produced the highest accuracy and performed better than the RF and the SVM with 0.94 OA, 0.89 KC, 0.96 P, 0.91 R and 0.93 F1.

Table 4. Model's performance of land-use binary mapping using S-2 dataset

No	Machine learning model	*OA*	*KC*	*P*	*R*	*F1*
1	Extreme Boosting (XGBoost)	0.94	0.89	0.96	0.91	0.93
2	Random Forests (RF)	0.92	0.85	0.88	0.87	0.91
3	Support Vector Machine (SVM)	0.86	0.79	0.84	0.82	0.85

The land use binary classification maps were created for the Wests and Cookies area using the XGBboost model using S-2 dataset and Google Earth imagery (Fig. 5). The classified map includes only bare soil and vegetation classes. Based on the binary classification maps, the precise locations belonging to the bare-soil pixels were used as a guide for sampling agricultural SOC field collection.

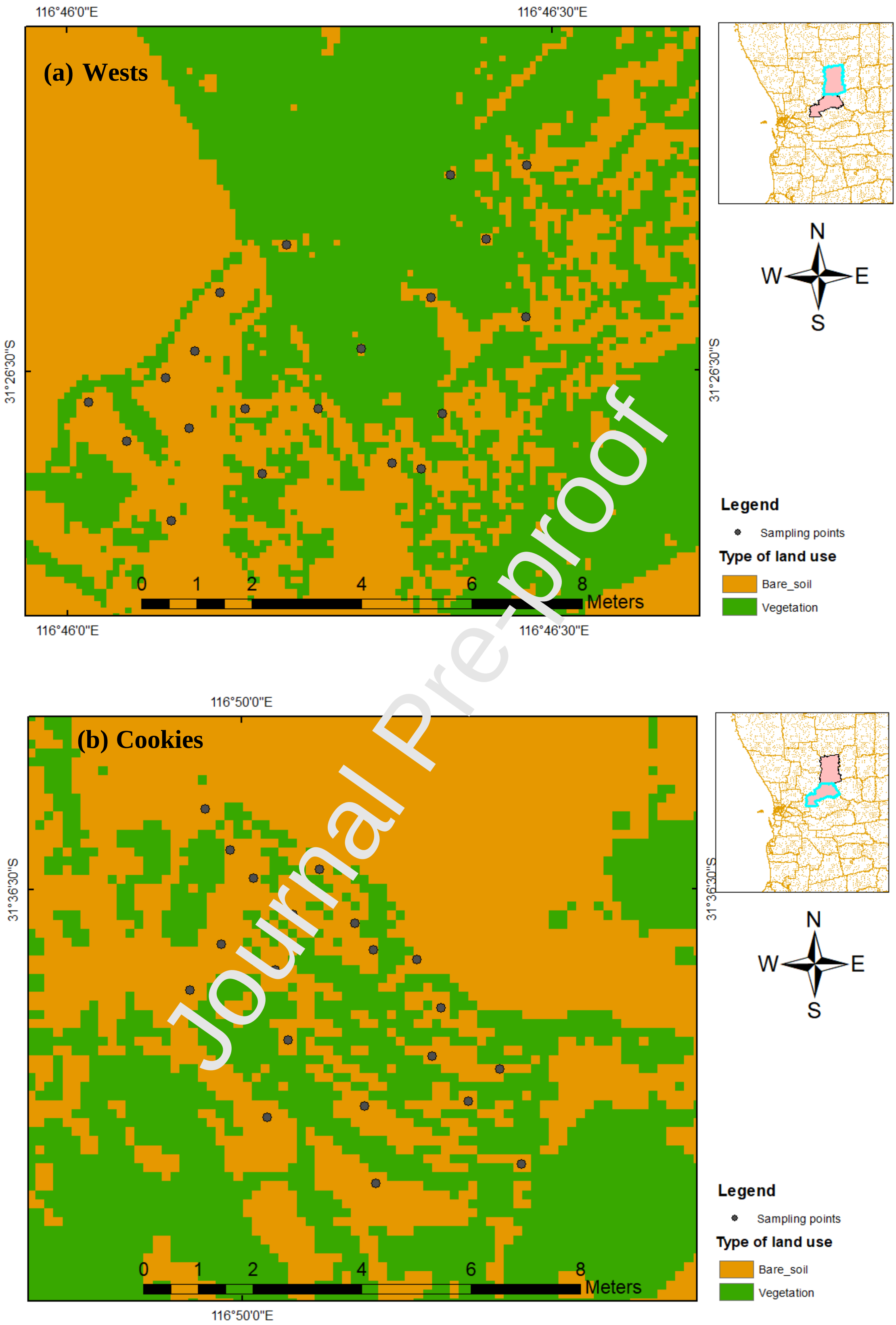


Figure 5. Land use binary classification map derived from the XGBoost model using S-2 and sampling points selection: (a) Wests, and (b) Cookies

The correlation coefficient between the input features derived from S2 data, VIs, and SIs with measured agricultural SOC was computed and illustrated in table 5. According to Table 5, the Ratio Vegetation Index (RVI), the Normalized Difference Vegetation Index (NDVI), and the Soil Adjusted Vegation Index (SAVI) presented the highest correlation with measured agricultural SOC among 21 predictor features derived from the S-2 image. These indices revealed positive correlations with agricultural SOC. In contrast, the lowest correlations were observed between Brightness Index 2 (BI2) and agricultural SOC. Vegetation and Soil Indices confirmed a higher correlation with agricultual SOC than ten S-2 multispectral bands. While vegetation indices illustrated positive correlations with agricultural SOC, most soil indices including BI, CI, and RI demonstrated negative correlations.

Table 5. Pearson's correlation analysis of S-2 derived predictor indicators and measured SOC

S2_Bands_Index	Correlation coefficient	S2_VI_BI_Index	Correlation coefficient
B2	-0.056	RVI	0.409
B3	-0.043	NDVI	0.419
B4	-0.162	GNDVI	0.167
B5	-0.131	NDI45	0.116
B6	-0.011	SAVI	0.470
B7	0.059	MCARI	0.088
B8	0.125	IRECI	0.377

S2_Bands_Index	Correlation coefficient	S2_VI_BI_Index	Correlation coefficient
B8A	0.170	BI	-0.113
B11	-0.022	BI2	0.005
B12	-0.025	CI	-0.296
		RI	-0.059

Table 6 shows the Pearson's correlation analysis of S-1 derived predictor indicators and measured agricultural SOC. VV, (VV+VH)/2, VH_GLCM Mean, VH_GLCM Variance, VV_Dissimilarity, VV_Homogeneity, VV_Angular Second Moment, VV_Entropy, VV_GLCM Mean, VV_GLCM Variance demonstrated the highest correlation with agricultural SOC compared to other predictor features generated from S1 data. Most GLCM textures showed strong correlations with agricultural SOC content. Four out of five S-1 SAR transformation bands (VH-VV; VV-VH; VV/VH; and VH/VV) had weak relationships with agricultural SOC.

Table 6. Pearson's correlation analysis of S-1 derived predictor indicators and measured SOC

S1_Index	Correlation coefficient	S1_Index	Correlation coefficient	S1_Index	Correlation coefficient
VH	0.389	VH_Homogeneity	-0.100	VV_Dissimilarity	0.417
VV	0.433	VH_Angular Second Moment	-0.047	VV_Homogeneity	-0.416
(VH+VV)/2	0.439	VH_Energy	-0.083	VV_Angular Second Moment	-0.431
VH-VV	0.251	VH_Maximum Probability	-0.067	VV_Energy	-0.349
VV-VH	-0.243	VH_Entropy	0.106	VV_Maximum Probability	-0.363
VV/VH	-0.118	VH_GLCM Mean	0.434	VV_Entropy	0.432
VH/VV	0.118	VH_GLCM Variance	0.437	VV_GLCM Mean	0.476
VH_Contrast	0.243	VH_GLCM Correlation	-0.211	VV_GLCM Variance	0.468
VH_Dissimilarity	0.168	VV_Contrast	0.359	VV_GLCM Correlation	-0.328

3.2. Evaluation and comparison of scenarios and different ML models

Five scenarios with varied features generated from S-2 and S-1 sensor were tested using the XGBoost technique (Table 7). The SC5 with the best possible number of features derived from multi-sensor S-1 and S-2 produced the highest prediction accuracy compared to the others SCs. However, the SC3 with only seven predictor variables from S1 yielded the worst prediction performance. A combination of S-2 and S-1 derived predictor features showed the highest R^2 of 0.870 in the validation phase and the lowest RMSE of 1.818 tonC/ha.

Table 7. Model performance of the XGBoost technique in six scenarios

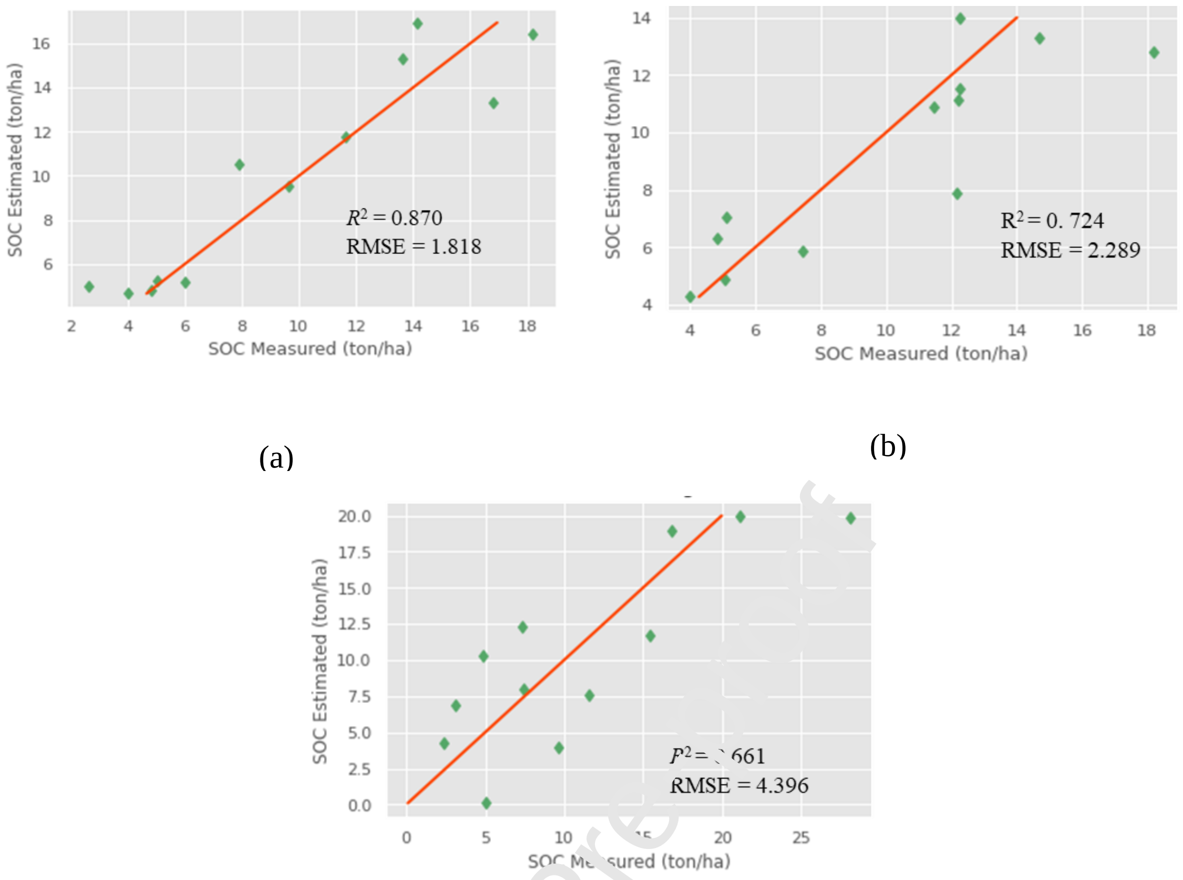
Scenario (SC)	Number of features	R^2 training (70%)	R^2 validation (30%)	RMSE (Ton C/ha)
SC1	10 features (10 S-2 bands only)	0.713	0.443	3.160
SC2	21 features (10 S-2 bands, 7 bands VIs, and 4 bands SIs)	0.891	0.625	2.370
SC3	7 features (2 bands from dual polarization, 5 SAR transformed bands)	0.559	0.254	3.004
SC4	27 features (2 bands from dual polarization, 5 SAR transformed bands, and 20 bands created from GLMC)	0.998	0.584	2.471
SC5	48 features (21 S-2 bands and 27 S1-bands)	0.927	0.870	1.818

To compare the effectiveness of the proposed XGBoost model using multi-source EO data fusion, two other well-known ML algorithms were selection for the comparison. The performance of the three ML algorithms on agricultural SOC retrievals are presented in Table 8. The SVM model performance in the agricultural SOC prediction was the lowest (R^2 = 0.661) and the RMSE value (4.396 ton/ha) was higher than those produced the XGBoost and the RF model. The XGBoost model with 48 predictor variables derived from a combination of S-2 image and S-1 image yielded the most accurate for agricultural SOC prediction in the validation phases (R^2 = 0.870, and RMSE = 1.818 ton/ha), followed by the RF model (R^2 = 0.724 and RMSE= 2.289 ton C/ha, and the SVM model (R^2 = 0.661 and RMSE= 4.396).

Table 8. Performance comparison of ML algorithms on agricultural SOC estimation

No	Machine learning model	R^2 training (70%)	R^2 testing (30%)	RMSE (Ton C/ha)
1	Extreme Boosting (XGBoost)	0.927	0.870	1.818
2	Random Forests (RF)	0.827	0.724	2.289
3	Support Vector Machine (SVM)	0.999	0.661	4.396

Figure 6 indicates the scatter plots of the estimated versus measured agricultural soil organic carbon using three well-known ML techniques in testing phase. The proposed ML models with auxiliary variables from S-2 multispectral imagery and S-1 SAR data can successfully estimate the agricultural SOC. The XGBoost is better at prediction than the RF and SVM.



(a)

(b)

(c)

Figure 6. Scatter diagrams of the measured SOC and estimated SOC by (a) XGBoost, (b) RF, (c) and SVM.

3.3. Spatial distribution patterns of agricultural SOC maps

Based on scenario 5, the spatial distribution of agricultural SOC maps generated for the Wests and Cookies areas using a combination of S1 and S2 datasets integrated by the XGBoost model are demonstrated in Fig. 7. The max, min, mean and standard deviation (SD) values of the predicted agricultural SOC were 15.899 ton C/ha, 5.42 ton C/ha, 6.936 ton C/ha, and 0.45 ton C/ha, respectively. The XGBoost model produced the low level of uncertainty and stable prediction capabilities with the low average value of SD.

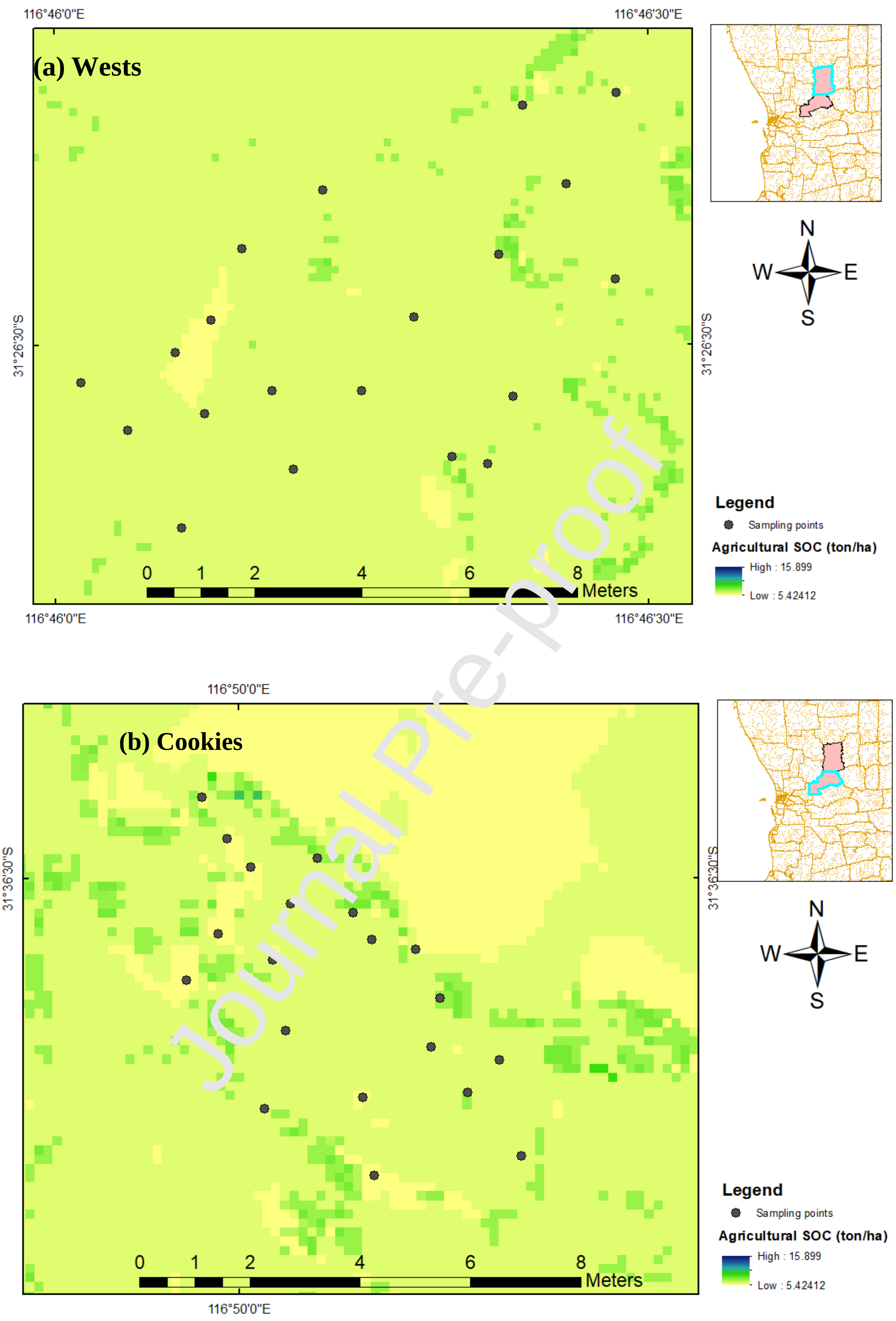


Figure 7. Spatial distribution characteristic of agricultural SOC in study areas: (a) Wests (a) and (b) Cookies using the proposed XGBoost combined data fusion.

4. Discussion

4.1. Performance of agricultural SOC prediction models

The prediction accuracy of agricultural SOC has been greatly influenced by the selection of predictor variables, ML algorithms, and level of data fusion (Table 7). The higher level of data fusion with more predictor features derived from Sentinel 2 and Sentinel 1 illustrated better prediction accuracy for retrieving agricultural SOC. This outcome is consistent with what Zhou et al (2020) and Castaldi et al (2019) reported. They indicated that the type of remote sensing data, predictor variables selection and the choice predictive models play important roles in SOC estimation (Castaldi et al., 2019). As well, combining S-2 and S-1 free-of-charge EO data can improve SOC prediction performance. Recent studies also stated that the multi-sensor data fusion has proved to be more effective than the single sensor approach in quantifying SOC for both mangrove SOC stocks and agricultural SOC content (Le et al., 2021; Zhou et al., 2020b).

The XGBoost predictive model is an efficient and effective gradient boosting algorithm which can be applied successfully for predictive modelling in SOC stocks research. The performance of the proposed XGBoost model combined with data fusion in the study performed well and outperformed the two well-known ML algorithms i.e. the RF and the SVM. The XGBoost algorithm is powerful and an advanced ML technique in predicting SOC stocks which is backed up in other recent studies (Ha et al., 2021; Ibrahem Ahmed Osman et al., 2021). The prediction results of the XGBoost in the study shows superior results (R^2 =0.87, RMSE = 1.818 tonC/ha) which are very much higher than the results of other studies noted in Table 1. The proposed framework using the 48 predictor features (10 multispectral bands, 7 vegetation indices, 4 soil indices, 2 bands from dual polarization, 5 SAR transformed bands, and 20 bands created from GLMC) derived from S1 and S2 combined with the XGBoost ML technique were powerful in agricultural SOC prediction. Importantly, the novel framework developed in this work is able to handle a small number of agricultural SOC samples, reflecting the robustness and cost-effectiveness of the model development for future and long-term agricultural SOC monitoring. However, more studies must be done on more sites, incorporating a wider geographical area.

4.2. Relative importance of predictor variables

The successful application of satellite RS images in predicting agricultural SOC has been proved in much research at the regional, national and global scale (Croft et al., 2012; Dvornikov et al., 2021; Hamzehpour et al., 2019; Mirzaee et al., 2016; Paul et al., 2020; Zhou et al., 2020a). However, most studies on this topic concentrated on mapping agricultural SOC based on optical imagery like S-2 imagery, which is due to the close relationship of Sentinel 2 derived indicators and SOC distribution. The present study illustrated that the predictor variables derived from both optical and SAR dataset are effective in estimating agricultural SOC. Similar observations were demonstrated by Yang and Guo (2019) (Yang & Guo, 2019). The relative importance of prediction features is presented in Fig. 8. Only 24 variables (10 features derived from S-2 and 14 features derived S-1) out of 48 variables were shown the high relative importance in the agricultural SOC.

Soil Adjusted Vegetation Index (SAVI) was identified as the most important predictor feature for agricultural SOC retrieval. It is due to its high sensitivity to soil characteristics (Huete 1988). The SAVI computed from the NIR and the Red bands also shows the strongest correlation coefficient (0.47) in Table 5, reflecting a high sensitivity to soil backgrounds and allowing to quantify the agricultural soil texture and SOC. The result is similar to the finding reported by Xue and Su (2017). The GLCM indicators, and dual polarization VV and VH derived from S-1 are also influential features. The contribution of the predictor variables computed from SAR data on determining agricultural SOC are more significant than S-2 derived variables. This is due to the capture ability of vegetation short-term variation characteristics of the Sentinel 1 sensor. Remarkably, the GLMC textures derived from

Sentinel 1 were not previously selected as the predictor features for agricultural SOC prediction. Nonetheless, it can be seen from Fig. 8 that GLMC bands from the VV polarization have been illustrated as being satisfactory predictor variables for estimating agricultural soil organic carbon. Future studies focusing on the SAR mechanism on agricultural SOC should be further investigated.

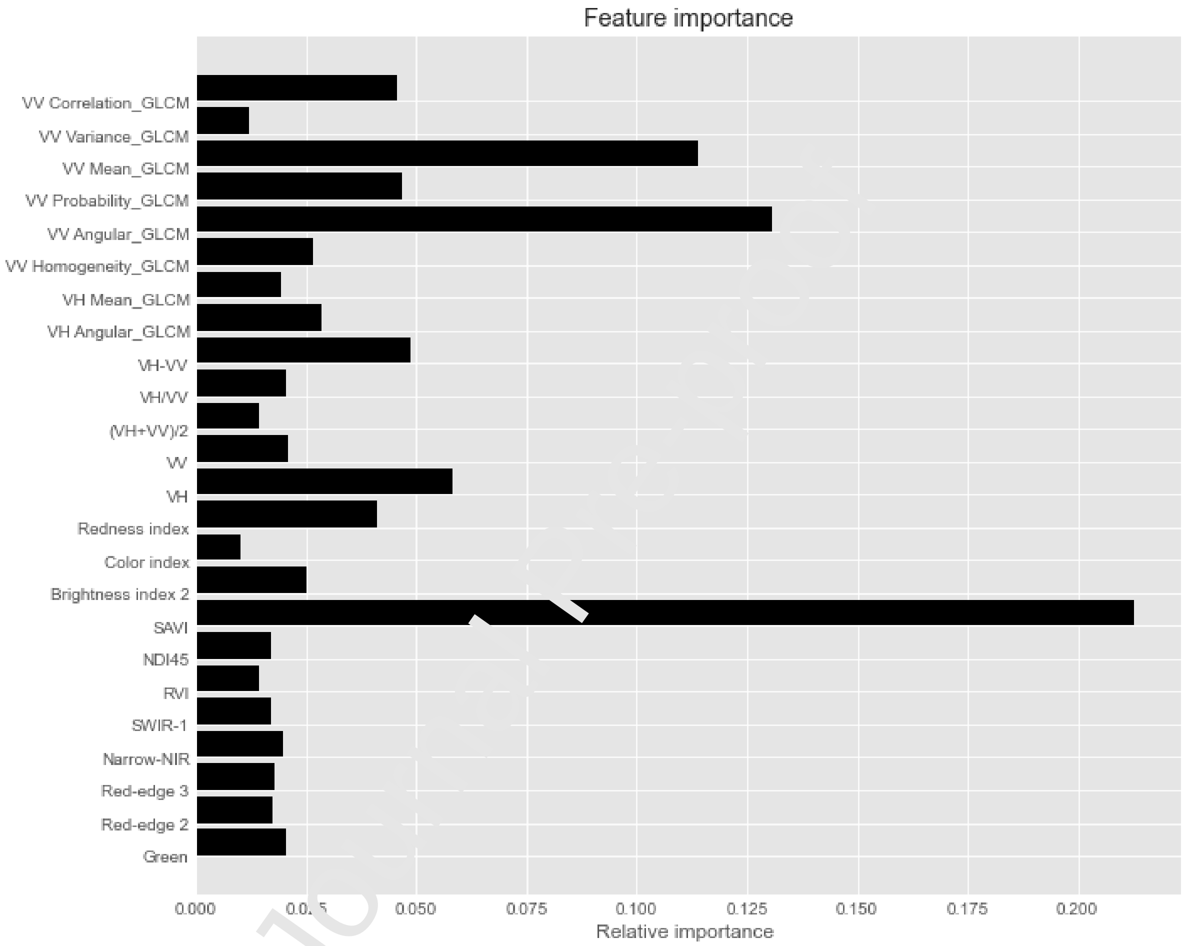


Figure 8. Variable importance of optimal features derived from multi-source EO data.

5. Conclusion

The present study pioneers the use of predictor features (dual polarizations and transformed bands) from SAR remote sensing imagery (S-1) and the fusion of predictor variables derived from optical remote sensing imagery (S-2) with a state-of-art machine learning technique (XGBoost). It is applied for predicting agricultural SOC in Western

Australia. Overall, the combination of S1 C-band dual polarimetric SAR and optical S2 datasets proved to be very useful for agricultural SOC prediction. High level of data fusion or multi-source sensor derived predictive variables illustrated significantly better prediction performance than a low level of data fusion or single sensor derived features. The proposed XGBoost model using multi-sensor data fusion demonstrated the highest prediction accuracy (R^2=0.870, RMSE= 1.818 ton/ha). In addition, the proposed model is able to derive agricultural SOC maps at 10m spatial resolution on regional scale with a precise accuracy. The binary land-use classification mapping using active learning to select bare soil sampling points and DPGS play important roles in the improvement of agricultural SOC prediction accuracy. Combining ensemble-based learning and active learning can enhance the estimates of agricultural SOC with only a small soil sample dataset. In short, this SOC prediction approach makes possible carbon neutrality for agriculture towards additional revenue via carbon credits.

Acknowledgements

This research were supported by the RIA Grant, University of Technology, Sydney, Australia (UTS, RIA NGO), Australian Postgraduate Research Intern (APR. Intern), and Astron Environmental Services company.

References

Aldana-Jague, E., Heckrath, G., Macdonald, A., van Wesemael, B., Van Oost, K. 2016. UAS-based soil carbon mapping using VIS-NIR (480–1000nm) multi-spectral imaging: Potential and limitations. *Geoderma*, **275**, 55-66.

Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., Bochtis, D. 2019. Remote Sensing Techniques for Soil Organic Carbon Estimation: A Review. *Remote Sensing*, **11**(6).

Breiman, L. 2001. Random Forests. *Machine Learning*, **45**(1), 5-32.

Castaldi, F., Hueni, A., Chabrillat, S., Ward, K., Buttafuoco, G., Bomans, B., Vreys, K., Brell, M., van Wesemael, B. 2019. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. *ISPRS Journal of Photogrammetry and Remote Sensing*, **147**, 267-282.

Castaldi, F., Palombo, A., Santini, F., Pascucci, S., Pignatti, S., Casa, R. 2016. Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon. *Remote Sensing of Environment*, **179**, 54-65.

Chen, T., Guestrin, C. 2016. XGBoost. in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794.

Chicco, D., Jurman, G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, **21(1)**, 6-6.

Cortes, C., Vapnik, V. 1995. Support-vector networks. *Machine Learning*, **20**(3), 273-297.

Cristianini, N., Ricci, E. 2008. Support Vector Machines. in: *Encyclopedia of Algorithms*, (Ed.) M.-Y. Kao, Springer US. Boston, MA, pp. 928-932.

Croft, H., Kuhn, N.J., Anderson, K. 2012. On the use of remote sensing techniques for monitoring spatio-temporal soil organic carbon dynamics in agricultural systems. *Catena*, **94**, 64-74.

Daughtry, C.S.T., Walthall, C.L., Kim, M.S., de Colstoun, E.B., McMurtrey, J.E. 2000. Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance. *Remote Sensing of Environment*, **74**(2), 229-239.

Delegido, J., Verrelst, J., Alonso, L., Moreno, J. 2011. Evaluation of Sentinel-2 red-edge bands for empirical estimation of green LAI and chlorophyll content. *Sensors (Basel, Switzerland)*, **11**(7), 7063-7081.

Dvornikov, Y.A., Vasenev, V.I., Romzaykina, O.N., Grigorieva, V.E., Litvinov, Y.A., Gorbov, S.N., Dolgikh, A.V., Korneykova, M.V., Gosse, D.D. 2021. Projecting the urbanization effect on soil organic carbon stocks in polar and steppe areas of European Russia by remote sensing. *Geoderma*, **399**.

Elhag, M., Bahrawi, J.A. 2017. Soil salinity mapping and hydrological drought indices assessment in arid environments based on remote sensing techniques. *Geosci. Instrum. Method. Data Syst.*, **6**(1), 149-158.

Escadafal, R. 1989. Remote sensing of arid soil surface color with Landsat thematic mapper. *Advances in Space Research*, **9**(1), 159-163.

Frampton, W.J., Dash, J., Watmough, G., Milton, E.J. 2013. Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. *ISPRS Journal of Photogrammetry and Remote Sensing*, **82**, 83-92

Forkuor, G., Hounkpatin, O.K.L., Welp, G., Thiel, M. 2017. High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. *PLOS ONE*, **12**(1), e0170478.

Gholizadeh, A., Žižala, D., Saberioon, M., Borůvka, L. 2018. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. *Remote Sensing of Environment*, **218**, 89-103.

Gitelson, A.A., Kaufman, Y.J., Merzlyak, M.N. 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment*, **58**(3), 289-298.

Gomez, C., Viscarra Rossel, R.A., McBratney, A.B. 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. *Geoderma*, **146**(3), 403-411.

Guo, L., Sun, X., Fu, P., Shi, T., Dang, L., Chen, Y., Linderman, M., Zhang, G., Zhang, Y., Jiang, Q., Zhang, H., Zeng, C. 2021. Mapping soil organic carbon stock by hyperspectral and time-series multispectral remote sensing images in low-relief agricultural areas. *Geoderma*, **398**.

Guo, Y., He, J., Li, S., Zheng, G., Wang, L. 2020. Evaluating the feasibility of GF-1 remote sensing comparison with hyperspectral data for soil organic carbon prediction and mapping. *IOP Conference Series: Earth and Environmental Science*, **545**, 012016.

Fu, X., Shao, M., Wei, X., Horton, R. 2010. Soil organic carbon and total nitrogen as affected by vegetation types in Northern Loess Plateau of China. *Geoderma*, **155**(1), 31-35.

Ha, N.T., Manley-Harris, M., Pham, T.D., Hawes, I. 2021. The use of radar and optical satellite imagery combined with advanced machine learning and metaheuristic optimization techniques to detect and quantify above ground biomass of intertidal seagrass in a New Zealand estuary. *International Journal of Remote Sensing*, **42**(12), 4712-4738.

Hamzehpour, N., Shafizadeh-Moghadam, H., Valavi, R. 2019. Exploring the driving forces and digital mapping of soil organic carbon using remote sensing and soil texture. *Catena*, **182**.

Heanes, D.L. 1984. Determination of total organic- C in soils by an improved chromic acid digestion and spectrophotometric procedure. *Communications in Soil Science and Plant Analysis*, **15**(10), 1191-1213.

Huete, A.R. 1988. A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, **25**(3), 295-309.

Ibrahem Ahmed Osman, A., Najah Ahmed, A., Chow, M.F., Feng Huang, Y., El-Shafie, A. 2021. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal*, **12**(2), 1545-1556.

Jin, X., Song, K., Du, J., Liu, H., Wen, Z. 2017. Comparison of different satellite bands and vegetation indices for estimation of soil organic matter based on simulated spectral configuration. *Agricultural and Forest Meteorology*, **244-245**, 57-71.

Khaleghi, B., Khamis, A., Karray, F.O., Razavi, S.N. 2013. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, **14**(1), 28-44.

Lal, R. 2008. Carbon sequestration. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **363**(1492), 815-830.

Le, N.N., Pham, T.D., Yokoya, N., Ha, N.T., Nguyen, T.T.T., Tran, T.D.T., Pham, T.D. 2021. Learning from multimodal and multisensor earth observation dataset for improving estimates of mangrove soil organic carbon in Vietnam. *International Journal of Remote Sensing*, **42**(18), 6866-6890.

Louis J, Debaecker V, Pflug B, Main-Knorn M, Bieniarz J, Mueller-Wilm U, et al. Sentinel-2 sen2cor: L2a processor for users. Proceedings of the Living Planet Symposium, Prague, Czech Republic, 2016, pp. 9-13.

Mathieu, R., Pouget, M., Cervelle, B., Escadafal, R. 1998. Relationships between Satellite-Based Radiometric Indices Simulated Using Laboratory Reflectance Data and Typic Soil Color of an Arid Environment. *Remote Sensing of Environment*, **66**(1), 17-28.

Mirzaee, S., Ghorbani-Dashtaki, S., Mohammadi, J., Asadi, H., Asadzadeh, F. 2016. Spatial variability of soil organic matter using remote sensing data. *CATENA*, **145**, 118-127.

Navarro-Pedreño, J., Almendro-Candel, M.B., Zorpas, A.A. 2021. The Increase of Soil Organic Matter Reduces Global Warming, Myth or Reality? *Sci*, **3**(1), 18.

Filipponi, F. 2019. Sentinel-1 GRD Preprocessing Workflow. *Proceedings*, **18**(1), 11.

Paul, S.S., Coops, N.C., Johnson, M.S., Krzic, M., Chandna, A., Smukler, S.M. 2020. Mapping soil organic carbon and clay using remote sensing to predict soil workability for enhanced climate change adaptation. *Geoderma*, **363**.

Petersen, E.H., Hoyle, F.C. 2016. Estimating the economic value of soil organic carbon for grains cropping systems in Western Australia. *Soil Research*, **54**(4), 383-396.

Pham, T.D., Yokoya, N., Nguyen, T.T.T., Le, N.N., Ha, N.T., Xia, J., Takeuchi, W., Pham, T.D. 2020. Improvement of Mangrove Soil Carbon Stocks Estimation in North

Vietnam Using Sentinel-2 Data and Machine Learning Approach. *GIScience & Remote Sensing*, **58**(1), 68-87.

Rouse, J., Haas, R.H., Schell, J.A., Deering, D. 1973. Monitoring vegetation systems in the great plains with ERTS.Salim, R.A., Islam, N. 2010. Exploring the impact of R&D and climate change on agricultural productivity growth: the case of Western Australia*. *Australian Journal of Agricultural and Resource Economics*, **54**(4), 561-582.

Six, J., Elliott, E.T., Paustian, K., Doran, J.W. 1998. Aggregation and Soil Organic Matter Accumulation in Cultivated and Native Grassland Soils. *Soil Science Society of America Journal*, **62**(5), 1367-1377.

Stevens, A., Udelhoven, T., Denis, A., Tychon, B., Lioy, R., Hoffmann, L., van Wesemael, B. 2010. Measuring soil organic carbon in croplands at regional scale using airborne imaging spectroscopy. *Geoderma*, **158**(1), 32-45.

Tucker, C.J. 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, **8**(2), 127-150.

Tuia, D., Volpi, M., Copa, L., Kanevski, M., Munoz-Mari, J. 2011. A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification. *IEEE Journal of Selected Topics in Signal Processing*, **5**(3), 606-617.

Vaudour, E., Gomez, C., Fouad, Y., Lagacherie, P. 2019. Sentinel-2 image capacities to predict common topsoil properties of temperate and Mediterranean agroecosystems. *Remote Sensing of Environment*, **223**, 21-33.

Venter, Z.S., Hawkins, H.J., Cramer, M.D., Mills, A.J. 2021. Mapping soil organic carbon stocks and trends with satellite-driven high resolution maps over South Africa. *Sci Total Environ*, **771**, 145384.

Vohland, M., Ludwig, M., Thiele-Bruhn, S., Ludwig, B. 2017. Quantification of Soil Properties with Hyperspectral Data: Selecting Spectral Variables with Different Methods to Improve Accuracies and Analyze Prediction Mechanisms. *Remote Sensing*, **9**(11).

Xue, J., Su, B. 2017. Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. *Journal of Sensors*, **2017**, 1353691.Yang, R.-M., Guo, W.-W. 2019. Modelling of soil organic carbon and bulk density in invaded coastal wetlands using Sentinel-1 imagery. *International Journal of Applied Earth Observation and Geoinformation*, **82**, 101906.

Zhou, T., Geng, Y., Chen, J., Liu, M., Haase, D., Lausch, A. 2020a. Mapping soil organic carbon content using multi-source remote sensing variables in the Heihe River Basin in China. *Ecological Indicators*, **114**.

Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D., Lausch, A. 2020b. High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. *Sci Total Environ*, **729**, 138244.

Zhou, T., Geng, Y., Ji, C., Xu, X., Wang, H., Pan, J., Bumberger, J., Haase, D., Lausch, A. 2021. Prediction of soil organic carbon and the C:N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. *Sci Total Environ*, **755**(Pt 2), 142661.