sustainability

Article

Using Machine Learning Algorithms Based on GF-6 and Google Earth Engine to Predict and Map the Spatial Distribution of Soil Organic Matter Content

Zhishan Ye [1], Ziheng Sheng [1], Xiaoyan Liu [1], Youhua Ma [1], Ruochen Wang [2], Shiwei Ding [1], Mengqian Liu [3], Zijie Li [4] and Qiang Wang [1,*]

[1] College of Resources and Environment, Anhui Agricultural University, Hefei 230036, China; yezhishan@stu.ahau.edu.cn (Z.Y.); shengzh@stu.ahau.edu.cn (Z.S.); liuxy@stu.ahau.edu.cn (X.L.); yhma@ahau.edu.cn (Y.M.); 18720454@stu.ahau.edu.cn (S.D.)
[2] Heinz College of Information Systems and Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213, USA; ruochenw@andrew.cmu.edu
[3] School of Plant Protection, Anhui Agricultural University, Hefei 230036, China; liumq@stu.ahau.edu.cn
[4] Realty Research Center, Nanjing Agricultural University, Nanjing 210095, China; zjli@stu.njau.edu.cn
* Correspondence: 28104@ahau.edu.cn

Abstract: The prediction of soil organic matter is important for measuring the soil's environmental quality and the degree of degradation. In this study, we combined China's GF-6 remote sensing data with the organic matter content data obtained from soil sampling points in the study area to predict soil organic matter content. To these data, we applied the random forest (RF), light gradient boosting machine (LightGBM), gradient boosting tree (GBDT), and extreme boosting machine (XGBoost) learning models. We used the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) to evaluate the prediction model. The results showed that XGBoost (R^2 = 0.634), LightGBM (R^2 = 0.627), and GBDT (R^2 = 0.591) had better accuracy and faster computing time than that of RF (R^2 = 0.551) during training. The regression model established by the XGBoost algorithm on the feature-optimized anthrosols dataset had the best accuracy, with an R^2 of 0.771. The inversion of soil organic matter content based on GF-6 data combined with the XGBoost model has good application potential.

Keywords: geospatial modeling; machine learning; predictive mapping; remote sensing inversion; soil organic matter

Citation: Ye, Z.; Sheng, Z.; Liu, X.; Ma, Y.; Wang, R.; Ding, S.; Liu, M.; Li, Z.; Wang, Q. Using Machine Learning Algorithms Based on GF-6 and Google Earth Engine to Predict and Map the Spatial Distribution of Soil Organic Matter Content. *Sustainability* **2021**, *13*, 14055. https://doi.org/10.3390/su132414055

Academic Editor: Jeroen Meersmans

Received: 15 November 2021
Accepted: 17 December 2021
Published: 20 December 2021

1. Introduction

Soil organic matter (SOM) is an important factor considered in soil surveys and environmental quality assessments [1–4], a key factor that participates in the global carbon cycle [5–7], and is an important indicator for judging the level of soil fertility [8,9]. Determining the SOM content is vital for achieving sustainable agricultural development [10] and ecological civilization [11], supporting ecosystem services [12] and improving crop productivity [13,14].

With the increasing demand for informatization for precision agriculture, accurately, quickly, and extensively estimating the SOM content has become challenging for many researchers. The traditional geostatistical method for predicting SOM content involves the measurement of SOM content in many samples, with the help of scale deductions and related geostatistical models [15–17]; the core theories are variogram and kriging interpolation [18]. Although the geostatistical method is reliable in theory, its main limitation in predicting the SOM content is that the sampling dataset has difficulties in meeting the stationarity assumption under complex terrain environments, and a large amount of data is needed to meet the needs of spatial autocorrelation. Collecting many soil samples in a research area inevitably leads to problems such as a long sampling period, high economic cost, and low prediction accuracy of a single data source [19,20]. With the development of high-resolution satellite remote sensing technology, researchers have found that different SOM contents have unique spectral response characteristics in the visible and infrared bands [21]. Researchers combine spectral characteristics with SOM content using remote sensing technology. Retrieving SOM content is, thus, a hot topic in soil science research [22–28].

The quantitative inversion of the reflectance spectra of SOM began in the 1960s [29]. Researchers have found a correlation between the spectral reflectance of the soil and SOM content. The laboratory-based prediction of SOM content can only provide point-scale prediction data [30–33], not landscape-scale prediction data. Satellite remote sensing can be used to produce real-time, dynamic, macroscopic, accurate, and low-cost predictions. The dynamic monitoring of SOM has received extensive attention from the soil science community [22–25]. One of the best options for the inversion of SOM content is considered to be establishing a quantitative SOM inversion model using ground-measured data and multi-spectral satellite remote sensing data with a high spatial resolution, spectral resolution, rich information, and high positioning accuracy to predict large-scale SOM content [26,27]. The SOM content of soil presented in a remote sensing image can be predicted by a regression model. This Model established a relationship between the spectral reflectance and SOM content of ground samples. When putting the spectral reflectance information of non-sampling points into it, the SOM content is the result, which was obtained from the calculation in the regression equation.

Since the 1970s, with the use of Landsat data, multispectral satellite data have been widely used in soil surveys [34–37]. Researchers have begun to combine satellite remote sensing data and soil data to establish a regression relationship between soil band reflectivity and organic matter content to predict large-scale SOM content. Different regression algorithms and mathematical transformations of reflectivity [38–40] affect the accuracy of organic matter prediction. The SOM inversion method is not universal for all regions and performs differently in different practical applications. Therefore, remote-sensing inversion of SOM is another research hotspot [41–43]. Machine learning algorithms have been gradually introduced for the prediction of various soil properties in the fields of mathematics and computers [44–47]. Numerous experiments have proven that machine learning algorithms perform well in analyzing nonlinear SOM characteristics and are more effective for multi-source and multi-feature data [22,46–49]. Compared with traditional linear regression algorithms, machine learning algorithms have a higher prediction accuracy and faster running speed for SOM inversion. Support vector machines and random forest (RF) algorithms have been widely used in previous single-element SOM content inversions [22,46–50]. With the update and iteration of the machine learning framework, models with higher prediction accuracy and better performance, such as the light gradient boosting machine (LightGBM) [45,51], gradient boosting tree (GBDT) [52–54], and extreme gradient boosting machine (XGBoost) [44,55,56] are now being widely used in agriculture, although rarely for SOM content prediction and research.

In this study, we selected Hefei City, Anhui Province, China as the research area to explore the prediction accuracy of SOM based on multi-source data, such as multi-spectral data, elevation, slope, vegetation index, cultivated land planting situation, and soil type. Our aim was to find the most accurate, fastest, and most stable SOM content inversion method suitable for farmland soil in this research area, to provide a valuable reference for the spatial estimation of SOM content.

2. Materials and Methods

2.1. Overview of the Research Area

The study area is located in Hefei, the capital city of Anhui Province, in eastern China, and is the sub-central city of the Yangtze River Delta City Group (116°41′–117°58′ E, 30°57′–32°32′ N; Figure 1). It has a humid subtropical monsoon climate with an average annual temperature of 15.7 °C and an average annual rainfall of approximately 1000 mm.

The study area includes hills, low mountains, and low-lying plains. The entire area is dominated by hills between the Yangtze and Huai Rivers. The main soil types are anthrosols and luvisols, accounting for approximately 85% of all soil types. The soil profile has a good structure and high nutrient content.

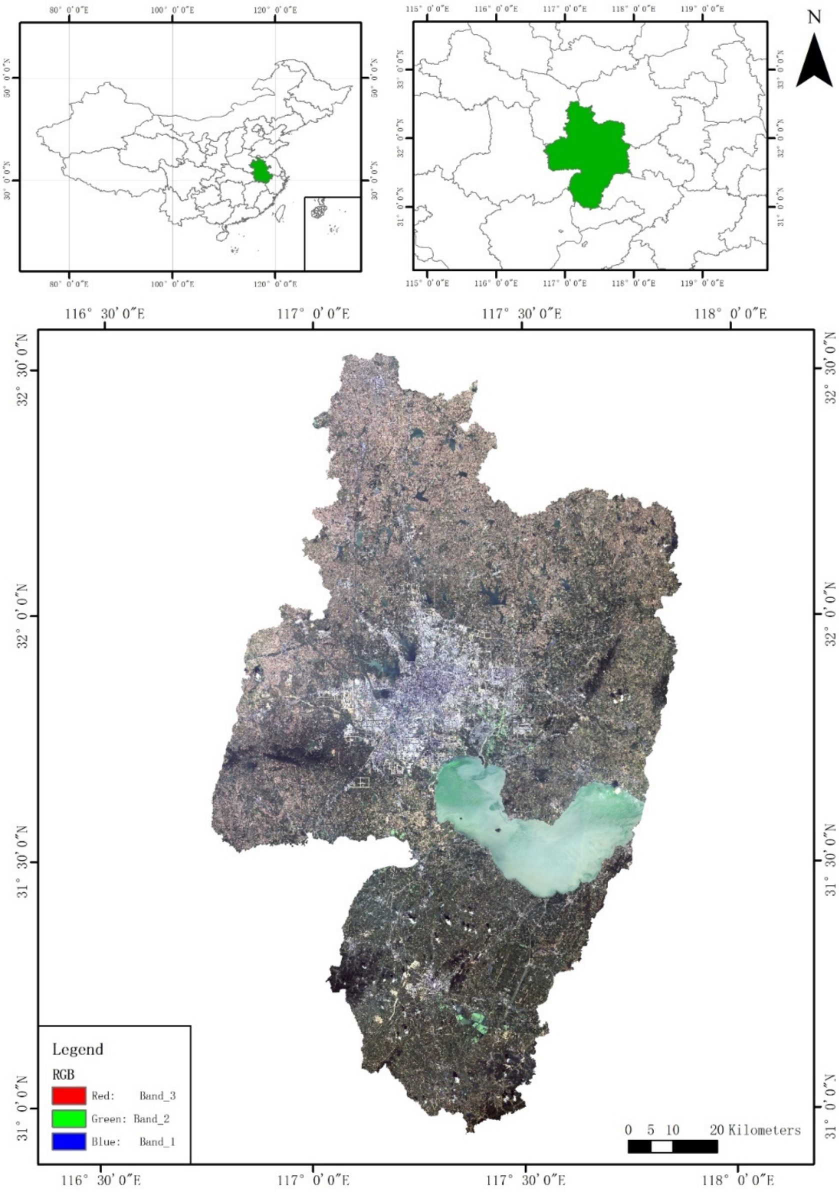

Figure 1. Location of the study area.

With a total area of 11,445 km^2 and an arable land area of 562,900 ha, Hefei is an important agricultural area in China. Cultivated land is mainly planted with grain crops, with the rotation of grain and non-food crops. The northern part of the study area is mainly planted with wheat, and the cultivated land in the western part of the study area uses a rotation of wheat and commercial forests.

2.2. Soil Sample Source

In October 2018, we sampled the cultivated soil in the study area, and a global positioning system (GPS) was used to record the coordinate information of the soil samples (Figure 2). According to the requirements of soil sampling point layout in DZ/T 0295–2016 Specification of Land Quality Geochemical Assessment, 295 topsoil samples were randomly arranged. We used soil auger to sample 0–20 cm soil column samples from the ground surface. About three to five subsoil columns were collected within a radius of 10 m around the sampling points to form one sample. The 295 soil samples were air-dried, ground, and subjected to other pretreatments, and the organic matter content of the soil was determined by the potassium dichromate-external heating method [57].

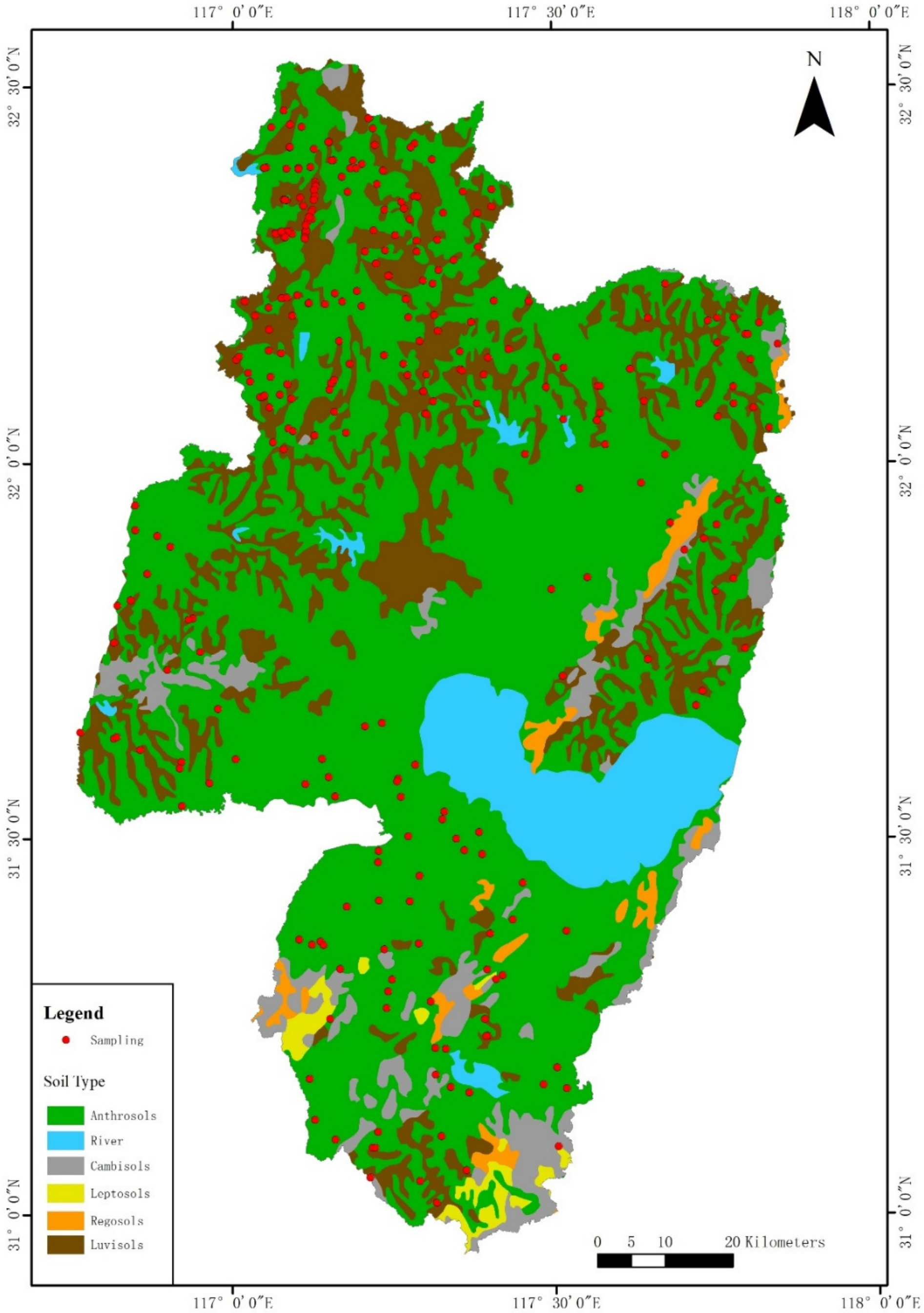

Figure 2. Distribution of sampling points on the soil type map.

2.3. Remote Sensing Data Processing

2.3.1. GF-6 Image Processing

Launched in June 2018, Gaofen-6 (GF-6) was China's first high-resolution satellite for precision agricultural observations. It has a high resolution, wide coverage, and high localization rate. In this study, we used GF-6 as the data source for obtaining multispectral (MUX) and panchromatic (PAN) data with less than 10% cloud cover at the beginning of October 2018. In ENVI 5.3, we processed the remote sensing image data by radiometric calibration and atmospheric correction, and the apparent reflectance values were converted into ground reflectance values. The reflectance values of four bands (B1, B2, B3, and B4) of the soil sample points were extracted using the ArcGIS 10.6 software platform (Table 1).

Table 1. Gaofen-6 (GF-6) band parameter information.

Band	Range (μm)
B1	0.45–0.52
B2	0.52–0.60
B3	0.63–0.69
B4	0.76–0.90

2.3.2. Annual Maximum Synthetic Data

The Google Earth Engine (GEE) platform is an online computing platform developed by Google for macro-scale remote sensing and analysis. Combining the massive remote sensing image data and powerful computing capabilities of the GEE platform, we examined the Landsat8 image dataset from January to December 2018 and from January to December 2019. On this basis, we used the GEE synthesis algorithm to obtain the following maximum synthetic vegetation indices [58] annually pixel by pixel in the study area: the normalized difference vegetation index (NDVI), ratio vegetation index (RVI), difference vegetation index (DVI), and 2018 normalized difference water index (NDWI). As the input feature of the model, we extracted the maximum synthetic vegetation index of the ground sample points.

2.3.3. Terrain Data

The 30 m resolution Shuttle Radar Topography Mission (SRTM) image data provided by the GEE platform were used to extract the sampling point elevation (digital elevation model (DEM)) and slope data [59].

2.4. Soil Type and Cultivated Land Planting Situation

The soil type data in the study area were obtained from the soil database of the Department of Agriculture and Rural Affairs of Anhui Province. The soil types in the study area are primarily anthrosols and luvisols. Cultivated land planting data were obtained from the third national land survey database of the Department of Natural Resources of Anhui Province. Cultivated land refers to dry land, paddy field, and irrigated land. The cultivated land-planting situation of the sample points in the study area is largely divided into planting food crops and rotation with non-food crops. Food crop in the study area refers to rice, wheat, and corn. Non-food crops in the study area refer to vegetables and fruits.

2.5. Spatial Distance Data

Human life and production activities have a long-term impact on the fertility of cultivated land [60,61]. The straight-line distance from the sampling point to the nearest residential area was calculated using the ArcGIS platform as a feature of SOM inversion (Table 2).

2.6. Model Building and Testing

As the ground truth data, we selected the soil samples collected in the study area in October 2018, and the GF-6 satellite image data from 4 October 2018 were the remote sensing data. We used Python-based random forest, LightGBM, GBDT, and XGBoost algorithms to establish an inversion model for SOM and remote sensing data. We used the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and runtime to evaluate the prediction model.

Table 2. Names and descriptions of the datasets used in this study.

Name of Dataset	Dataset	Resolution	Source
Soil reflectance data	B1-B4	8 m	GF-6
Google Earth Engine (GEE) maximum annual indices	2018NDWIx	8 m	Landsat8
	2018NDVIx	8 m	
	2019NDVIx	8 m	
	2018DVIx	8 m	
	2019DVIx	8 m	
	2018RVIx	8 m	
	2019RVIx	8 m	
GEE digital elevation model (DEM) data	DEM	30 m	Shuttle Radar Topography Mission (SRTM)
	Slope	30 m	
Soil and cultivated land planting situation	Soil type	Vector data	Department of agriculture
	cultivated land planting situation	Vector data	
Geostatistical data	Distance	Vector data	Department of natural resources

2.6.1. Model Introduction

Python-based machine learning models.

1. RF model: The RF algorithm is an ensemble learning method proposed by Breiman in 2001 [62]. The model is a bagged algorithm that contains multiple decision trees. The performance of a single regression tree is improved by combining multiple decision trees. The model output is the result of the integration of multiple decision trees.
2. GBDT model: The GBDT model is a combination of decision tree and boosting algorithms and was proposed by Friedman in 2001 [63]. The model is an integrated tree model that calculates the residuals between the actual and predicted values. The integrated algorithm model uses gradient, boosting, and decision tree to solve the classification problem and perform the regression prediction. Boosting refers to the offline combination of multiple weak classifiers to achieve a strong classifier, and gradient refers to the increase in flexibility and convenience when the model solves the loss function. Compared to the support vector machine model, the GBDT algorithm has fewer model parameters, faster calculation speed, and higher stability.
3. LightGBM model: As a part of the GBDT algorithm framework, LightGBM [64] internally integrates multiple decision trees and can integrate the decision results of multiple decision trees, avoiding the low accuracy shortcoming of the use of a single learning machine. The LightGBM algorithm adopts a leaf-wise growth strategy based on histograms, depth limitations, and exclusive feature bundling to increase the speed of calculation and improve training efficiency.
4. XGBoost model: Chen and Guestrin proposed a new machine learning algorithm called the XGBoost algorithm in 2016 [65]. It has achieved excellent results in many international data mining competitions, and its performance exceeds that of deep learning algorithms [66,67]. The XGBoost algorithm improved on the GBDT algorithm. The loss function is determined by a second-order Taylor expansion, and the regularization concept of the loss function is introduced. The number of constrained nodes and outputs are added to the loss function, which makes the XGBoost algorithm more accurate than the GBDT algorithm, and the algorithm is hard to overfit.

2.6.2. Model Evaluation and Tuning the Hyper-Parameters

The sampling points were divided into 265 points (90%) for the training set and 30 points (10%) for the test set for modeling and testing. R^2, RMSE, and MAE were used to evaluate the accuracy and stability of the model. K-fold cross-validation was used to validate the training set of the model, and the hyperparameters were adjusted using a grid search [68,69]. The K-fold cross-validation [70,71] divided the training set into K equal parts, leaving one part as the test set and the rest as the training set. The cross-validation was repeated K times, each sub-sample was verified once, and the average of the K results was taken as the model result. In this study, K = 10 means that we performed 10-fold cross-validation (Figure 3).

$$R^2 = \frac{\left[\sum_{i=1}^{n}(C_{ti} - \overline{C}_t)\cdot(C_{pi} - \overline{C}_p)\right]^2}{\sum_{i=1}^{n}(C_{ti} - \overline{C}_t)^2\cdot\sum_{i=1}^{n}(C_{pi} - \overline{C}_p)^2} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(C_{ti} - C_{pi})^2}{n}} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|C_{ti} - C_{pi}| \tag{3}$$

where n is the number of samples, C_{ti} is the true SOM content at the i-th sample point (g/kg), C_{pi} is the predicted SOM content at the i-th sample point (g/kg), $\overline{C}_t$ is the prediction of a true value, and $\overline{C}_p$ is the predicted value.

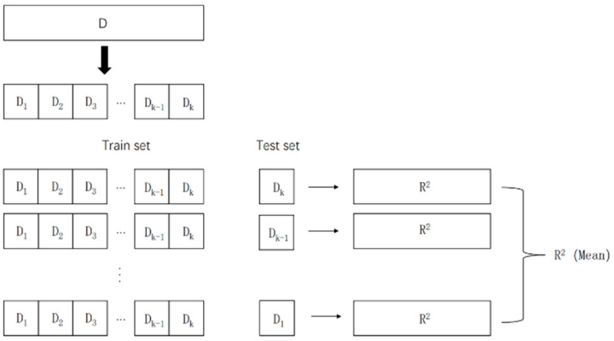

Figure 3. K-fold cross validation.

3. Results

3.1. *Statistical Analysis of Soil Organic Matter Content*

The soil types used in this study are listed in Table 3. Of the 295 total sampling points, 204 samples were taken from anthrosols (54.86% of the area) and 84 samples were taken from the luvisols (13.90% of the area). The other soil types covered smaller areas and consequently fewer sample points, without suitable conditions for modeling.

Table 3. Soil organic matter (SOM) content (g/kg) statistics for all samples, anthrosols samples, and luvisols samples.

Sampling Dataset	N	SOM						
		Max	Min	Mean	Standard Deviation	Kurtosis	Skewness	Coefficient of Variation
Whole sampling	295	44.6	9.8	23.19	5.894	0.384	0.048	0.254
Anthrosols	204	44.6	9.8	23.45	6.023	0.578	0.145	0.257
Luvisols	84	33.5	9.9	22.52	5.690	−0.477	−0.230	0.253

N = number.

Anthrosols is a type of soil that is strongly influenced by human cultivation, with a high organic matter content [72], with an average of 23.45 g/kg (Figure 4). Luvisols is leached soil developed on the parent material of Pleistocene loess in the fourth quarter [73], with an average organic matter content of 22.52 g/kg. Under the influence of different cultivated land planting situations, we found that the average SOM content of the crop area was 23.58 g/kg, which is greater than that of the crop rotation area at 22.60 g/kg (Figure 5). The *p*-value of the total sample dataset is 0.0077, which approaches zero, and the dataset tends to be normally distributed.

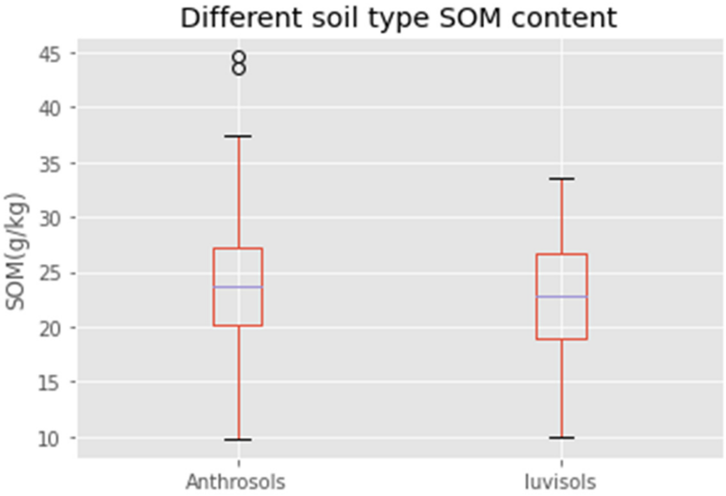

Figure 4. Box plot of SOM content in different soil types.

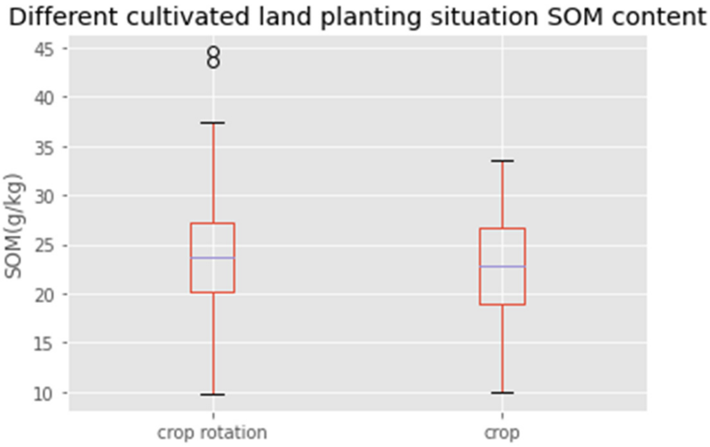

Figure 5. Box plot of SOM content in different cultivated land planting situations.

3.2. Correlation Analysis

As the features to establish a correlation heat map with SOM content, we used the GF-6 satellite's four-band reflectance data (B1–B4), six synthetic maximum vegetation indices in 2018 and 2019 (2018NDVIx-2019RVIx), 2018 synthetic maximum normalized water index data (2018NDWIx), DEM and slope data, land use data (cultivated land planting situation and soil type), and the spatial distance data between sampling points and residential areas (Figure 6). We concluded that the SOM content of the sampling point is highly correlated with the reflectance data of the four GF-6 bands.

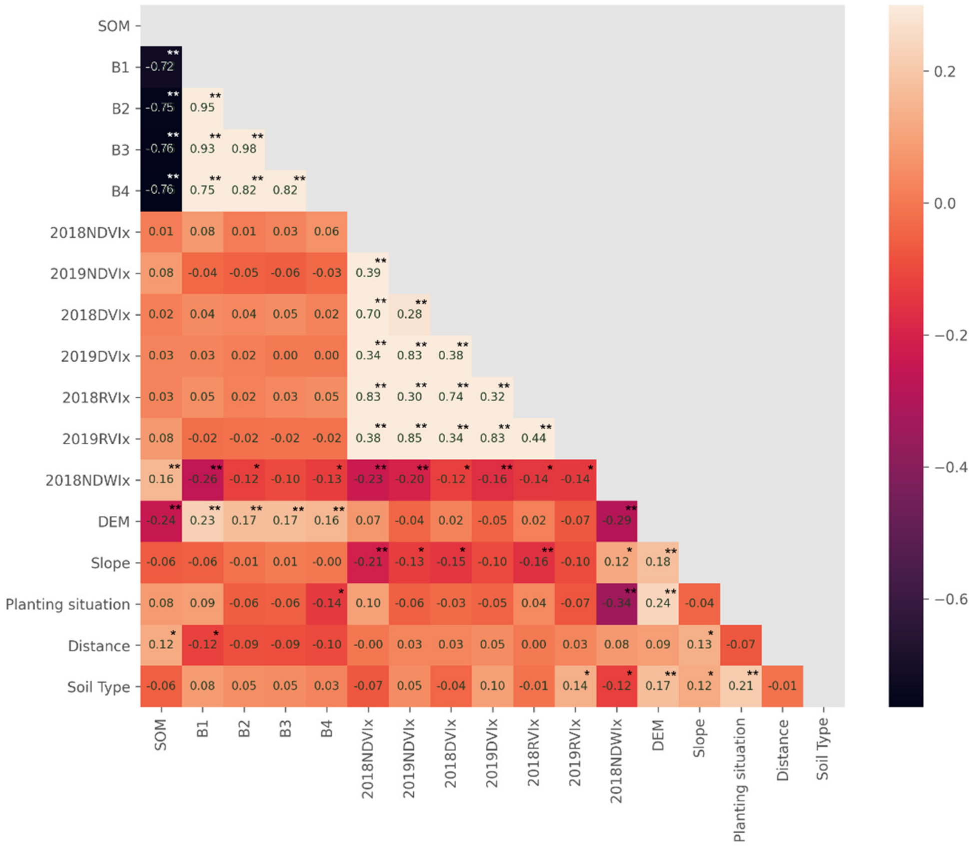

Figure 6. Correlation heat map between SOM and features of the study area (* $p \leq 0.05$; ** $p < 0.01$).

3.3. Prediction Results of Surface Soil Organic Matter Content

3.3.1. Model Hyperparameter Selection

We built the RF, GBDT, LightGBM, and XGBoost models in Python. A 10-fold cross-validation mean is used to evaluate the training set of the model. After using the default hyperparameters as inputs to obtain the initial results, we adjusted and optimized the corresponding hyperparameters of each model using GridSearchCV [74] to improve the accuracy of the test set of the model (Table 4).

Table 4. The optimal parameter, the default parameters, and the range of the grid search in each model.

Regression Model	Hyperparameters	Optimal Parameter	Default Parameters	Range of Grid Search
Extreme gradient boosting machine (XGBoost)	random_state	0	0	0
	n_estimators	15	100	0~100
	max_depth	2	6	1~10
	learning_rate	0.38	0.3	0.01~1.00
	min_child_weight	4	1	1~10
	gamma	0.1	0	0~1.0
Light gradient boosting machine (LightGBM)	random_state	0	None	0
	n_estimators	26	100	0~100
	max_depth	7	−1	1~10
	learning_rate	0.1	0.1	0.01~1.00
	subsample	0.1	1	0~1.0
Gradient boosting tree (GBDT)	random_state	0	None	0
	n_estimators	21	100	0~100
	max_depth	4	3	1~10
	learning_rate	0.18	0.1	0.01~1.00

Table 4. *Cont.*

Regression Model	Hyperparameters	Optimal Parameter	Default Parameters	Range of Grid Search
Random Forest (RF)	random_state	0	None	0
	n_estimators	83	100	0~100
	max_depth	8	None	1~10
	min_samples_split	9	2	1~10
	min_samples_leaf	1	1	1~10

3.3.2. Evaluation of Model Calculation Speed

Our experimental computer is based on Windows 10 system, Core i7 10710 processor, 16G RAM. Different machine learning models exhibit different computing speeds. We compared the calculation speed of different models by comparing the calculation completion time of different models on the same computer platform. We found that RF model was the slowest of the four machine learning models, and the LightGBM model was the fastest computing model (Table 5).

Table 5. Runtime of each model.

Regression Model	Runtime (s)
XGBoost	0.2
LightGBM	0.1
GBDT	0.3
RF	1.4

3.3.3. Model Accuracy Comparison

We used different algorithms to build models for all 16 features, and we evaluated the accuracy of the models by calculating the R^2, RMSE, and MAE values of the training and test sets (Table 6). The XGBoost model had the highest prediction accuracy (R^2 = 0.634, RMSE = 3.250, and MAE = 2.637), and the RF model had the lowest accuracy (R^2 = 0.551, RMSE = 3.591, and MAE = 2.698).

Table 6. Model prediction accuracy.

Regression Model	Performance Indicator		
	Coefficient of Determination (R^2)	Root Mean Square Error (RMSE)	Mean Absolute Error (MAE)
XGBoost	0.634	3.250	2.637
LightGBM	0.627	3.278	2.618
GBDT	0.591	3.432	2.780
RF	0.551	3.591	2.698

3.4. SOM Prediction Results of Different Datasets

3.4.1. Anthrosols Prediction Result

To explore the role of the different datasets in SOM prediction, samples of anthrosols were selected from the total dataset to form a anthrosols dataset. A 10-fold cross-validation was used to adjust the hyperparameters in the grid search of the anthrosols dataset (Table 7).

Table 7. Hyperparameter values for each model in the anthrosols dataset.

Regression Model	Hyperparameters	Optimal Parameter
XGBoost	random_state	0
	n_estimators	14
	max_depth	2
LightGBM	random_state	0
	n_estimators	60
	max_depth	6
	learning_rate	0.08
	subsample	0.01
GBDT	random_state	0
	n_estimators	27
	max_depth	3
	learning_rate	0.11
RF	random_state	0
	n_estimators	89
	max_depth	9
	min_samples_split	3
	min_samples_leaf	1

The XGBoost model(see Appendix A) best fits the anthrosols dataset (Table 8). The R^2 of the XGBoost model for anthrosols (0.748) was larger than that for the total dataset (0.634). Among the four machine learning models, the XGBoost, LightGBM, and GBDT models had the same operation speed, and the RF model had the slowest speed. Overall, the model had a faster calculation speed due to the smaller sample size.

Table 8. Prediction accuracy and runtime of each model for the anthrosols dataset.

Regression Model	Runtimes (s)	Performance Indicator		
		R^2	RMSE	MAE
XGBoost	0.2	0.748	1.858	1.405
LightGBM	0.2	0.355	2.974	2.198
GBDT	0.2	0.711	1.990	1.583
RF	1.4	0.741	1.880	1.649

3.4.2. Feature Selection for Anthrosols Dataset

Feature importance is an important reference when selecting features. The XGBoost model uses the gain criterion to calculate the importance of each feature when participating in model training. The gain is calculated by the contribution of the feature to each tree, that is, the contribution of each feature to the generative model. The higher the value, the greater the importance of this feature to the prediction of the model [75]. The importance of each feature in the anthrosols dataset on the XGBoost model was ranked to explore the influence of different features in the same dataset on the prediction accuracy of the model (Figure 7).

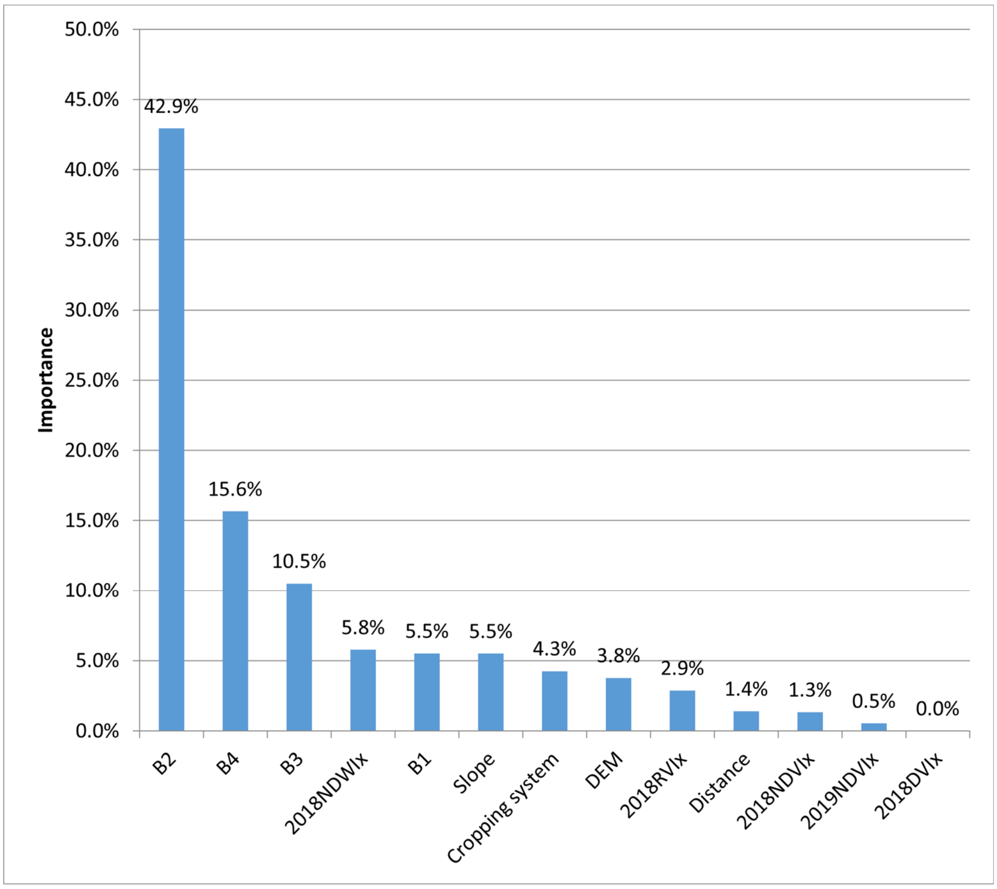

Figure 7. Feature importance ranking of the XGBoost model; the sum of importance of all features is 100%.

A 10-fold cross-validation mean was used to evaluate the influence of different numbers of features on the predictive ability of the model. From the experiment, we found that when the number of features involved in the fitting is nine, the maximum cross-validation mean is 0.565. Therefore, we took the top nine features with the most important features to participate in the training, and the model has the best fit (Figure 8).

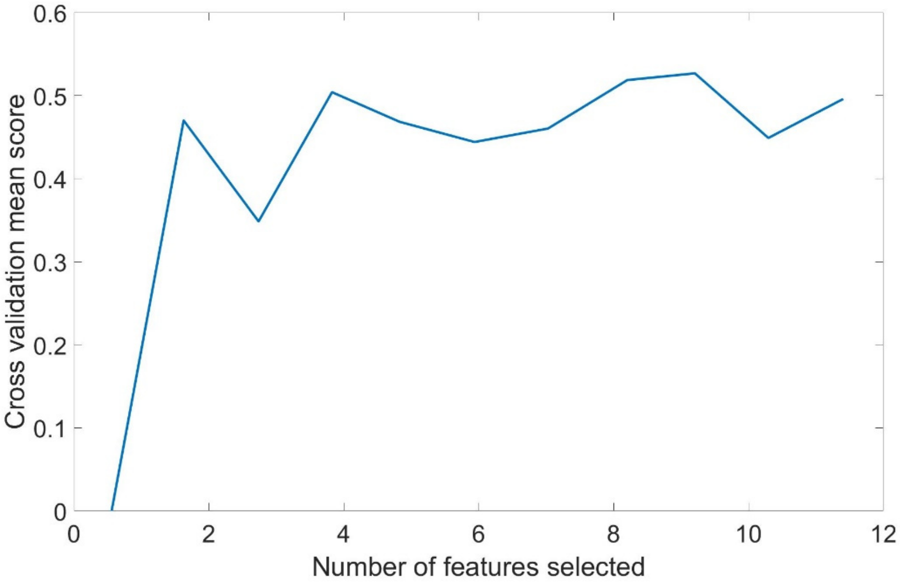

Figure 8. Influence of the number of features on the model cross-validation mean.

We selected the top nine features that had the most impact on the model, and we adjusted the hyperparameters of the XGBoost model through a grid search (random_state = 0, n_estimators = 20, and max_depth = 4).

Based on the prediction accuracy and efficiency of the four machine learning models, we found that the XGBoost model has a high accuracy, short runtime, and few hyperparameters that need to be adjusted, making it the most suitable for predicting the spatial distribution of SOM in the research area. After reducing the number of features, the predictive ability of the model was further improved, and R^2 reached 0.771. Thus, the predictive ability of the model can be improved by optimizing the number of features (Table 9).

Table 9. Anthrosols prediction accuracy and runtime of XGBoost model after feature optimization.

Regression Model	Runtime (s)	Performance Indicator		
		R^2	RMSE	MAE
XGBoost	0.2	0.771	1.773	1.474

3.5. Simulation of the Spatial Distribution of SOM

The XGBoost model trained on the anthrosols dataset after feature selection had the best prediction effect; therefore, we used this model to predict the spatial distribution of SOM content in the study area. Using Python to write a program on the Visual Studio Code platform, the data of the nine features per pixel were transferred to the trained model, and the pixel-by-pixel organic matter content was output to obtain the spatial distribution of the SOM content in the study area (Figure 9).

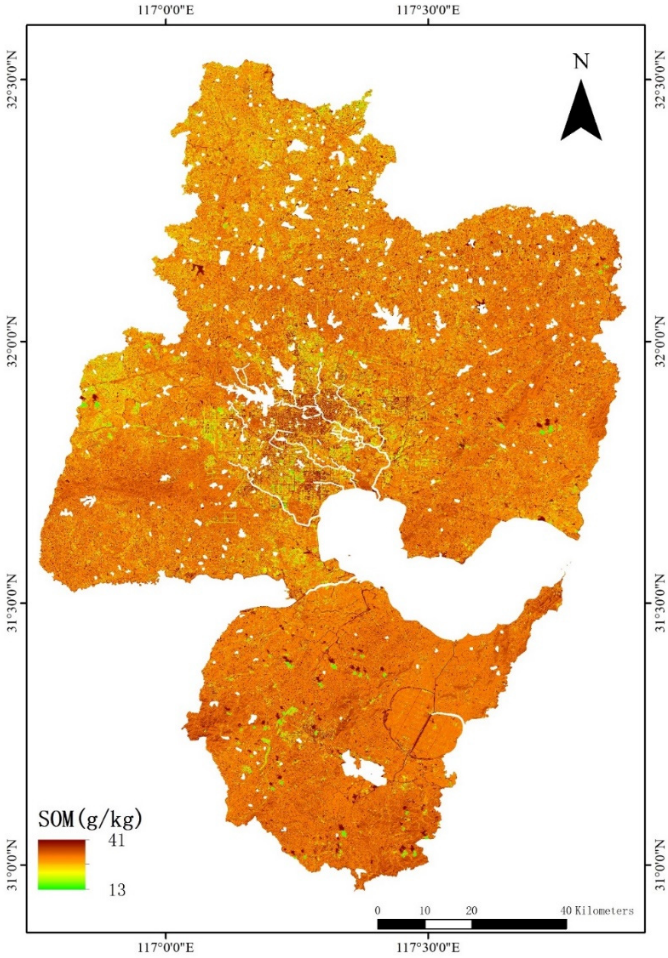

Figure 9. Prediction mapping of SOM content in the study area.

4. Discussion

4.1. Predictive Performance Analysis

In this study, we separately trained four machine learning models. From the observed performance of the model and the results of the training time, most hyperparameters need to be adjusted for the RF model, and the model's prediction accuracy and training time performance are the worst. LightGBM has the fastest training time, but it has a certain dataset, and its stability is low. XGBoost improves the GBDT model at the algorithm level, and its prediction accuracy and efficiency are higher than those of the GBDT model. As such, we found that the XGBoost model performs the best of the considered models.

The overall anthrosols prediction performance results indicate that the differences in the soil type impacts the prediction performance of the model. The reasons for this impact are as follows: (1) Anthrosols are the most widespread in the study area. The soil fertility is greatly affected by humans. The study area began to implement formula fertilization by soil testing policy in 2006. Government organized researchers to fertilize the cultivated land in the study area by measuring the physical and chemical composition of the soil (http://nync.ah.gov.cn/ (accessed on 8 December 2021)). Due to the high proportion of anthrosols cultivated land in the study area (54.86% of the area), regular water and fertilizer adjustments make anthrosols more fertile than other soil types, and the spatial variability of the SOM content is low [76,77]. (2) In a similar study, researchers found that different land use types have different prediction accuracies of SOM content. By comparing the prediction accuracy of dryland and paddy fields, they proposed that the higher coefficient of variation of the dryland data set and the coefficient of variation of the paddy field data set may lead to better prediction accuracy of the dryland data set [78]. In this study, the coefficient of variation for the organic matter content of the anthrosols dataset was greater than that of the total data set. An increase in the coefficient of variation may improve the prediction accuracy of the model.

4.2. Selection and Optimization Analysis of Multivariate Data Features

The total dataset used in this study has 16 features, and the nine features that were the most important to the model were selected in the final model feature optimization stage. The reflectance data of four bands of GF-6 (B1–B4), three annual maximum vegetation indices synthesized by GEE (2019RVIx, 2018RVIx, 2019DVIx), cultivated land planting situation data, and DEM data were analyzed. The two largest vegetation indices in 2019 significantly influenced the prediction results of the model. The maximum vegetation index synthesized by GEE may be useful in future soil attribute prediction studies. The cultivated land planting situations were divided into planting food crops and the rotation of food and non-food crops. The average SOM content of cultivated land for food crops is higher than that of cultivated land for rotation, and there are fewer abnormal values. This may be due to the organic fertilizer replacement policy implemented in the study area in 2016 (http://nync.ah.gov.cn/ (accessed on 8 December 2021)). Farmers in the study area are required to use organic fertilizers instead of part of the chemical fertilizers applied to the cultivated land where food crops are grown. This policy started in 2016, replacing chemical fertilizers with organic fertilizers, which can increase the SOM content [79]. Additionally, the altitude of the soil has a substantial impact on SOM, the correlation between elevation and SOM is -0.24, and the effect of elevation on SOM has also been confirmed in previous studies [80,81].

4.3. GF-6 Modeling Advantage

Researchers primarily use Landsat [21,27,36–38,58,82] and sentinel satellites [20,22,47,68,78] for the inversion of SOM content. Compared to the above multispectral and hyperspectral satellites, the use of GF-6 as the remote sensing image data source in this study has the following advantages: (1) Multispectral satellite data are more suitable for the inversion of SOM content than hyperspectral satellite data. Due to the large number of reflection bands of hyperspectral data, it is necessary to eliminate redundant and cluttered bands with low correlation before establishing a regression model [50], and (2) GF-6 was the first high-resolution satellite developed by China for precision agriculture. Using GF-6 to predict the distribution of SOM content provides a basis for the development of precision agriculture policies. The province where the study area is located has been actively carrying out the construction of well-facilitated farmland [83] since 2015, and clarifying the SOM content of the cultivated land is of great significance to the construction of well-facilitated farmland (http://nync.ah.gov.cn/ (accessed on 8 December 2021)). Well-facilitated farmland refers to farmland with leveled land, concentrated contiguous areas, complete facilities, supporting farmland, fertile soil, good ecology, strong disaster resistance, is compatible with modern agricultural production and management methods that are suitable for droughts and floods, and is a high and stable yield farmland. The results of this study can be used to monitor the SOM content of farmland when constructing and evaluating well-facilitated farmland.

4.4. Model Efficiency Analysis

Traditional SOM inversion methods typically use single-reflectivity mathematical transformations [38,41,50], such as reciprocal transformation, exponential transformation, and log transformation, as the inputs of the model. Such models have low accuracy and slow computational speed [38,50]. The RF machine learning model is unsuitable for future large-scale organic matter inversion research, and XGBoost models will gradually become mainstream in the future of soil property inversion.

4.5. Limitations of the Study

The SOM prediction based on a single soil type has limitations, which are mainly related to the soil type and topography of the specific research area. The soil sampling and remote sensing image selection in this study are all after the autumn harvest in this study area. The SOM prediction study was carried out during the bare soil period of the cultivated land, and there was no vegetation affecting the spectral reflectance of the soil. Further work is needed to prove whether this model can be used for SOM prediction in different regions and at different time periods.

4.6. Model Selection

Our research used the tree-based algorithm of machine learning. Tree-based algorithm models perform well in predicting linear problems, and researchers can easily interpret the model's predictions from the perspective of input features [44–47]. Deep learning models perform very well in the research of large data sets [84,85], but it is very difficult for researchers to explain the performance of the models in terms of their input features and model parameters [86]. In future research, we will try to use deep learning models to solve the problem of predicting the spatial distribution of SOM content.

5. Conclusions

In this study, we used GF-6 satellite, terrain, and soil type data and combined these data with actual ground measurement data to build a remote sensing monitoring method for SOM content based on XGBoost machine learning. We drew the following conclusions:

(1) By comparing the R^2, RMSE, and MAE values of each model, the XGBoost model was found to be the most suitable for predicting the spatial distribution of SOM in the study area. The R^2, RMSE, and MAE values of the XGBoost model based on the optimized anthrosols dataset were 0.771, 1.773, and 1.474, respectively.

(2) In terms of operating efficiency, the run times of the XGBoost, LightGBM, and GBDT models were shorter than those of the traditional RF model.

(3) Machine learning methods such as XGBoost can achieve rapid and economical inversion of SOM content, allowing their application in precision agriculture.

Author Contributions: Resources, Y.M. and Z.L.; software, Z.Y., R.W. and S.D.; visualization, Z.S., M.L. and X.L.; writing, Z.Y. and M.L.; supervision, Q.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China: (Grant number: 41801234), and the Collaborative Innovation Project of Anhui Provincial Department of Education (No. GXXT-2019-047).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Google Earth Engine (GEE) maximum annual indices and DEM data are publicly available online (https://earthengine.google.com/ (accessed on 8 December 2021)). Soil sample data were obtained from Department of Agriculture and Rural Affairs of Anhui Province.

Acknowledgments: We thank the National Natural Science Foundation of China: (Grant number: 41801234), and the Collaborative Innovation Project of Anhui Provincial Department of Education (No. GXXT-2019-047) for their support. We are also grateful to the editor and the reviewers for their helpful comments.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

XGBoost prediction model based on the anthrosols dataset.

```
import pandas as pd
import xgboost as xgb
import numpy as np
from sklearn.model_selection import cross_val_score
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import OneHotEncoder
from sklearn.model_selection import train_test_split

data = pd.read_csv(‘paddy.csv’)

#one-hot enconder
x_onehot = data[‘soil’]
values = np.array(x_onehot)
x_one_hot = OneHotEncoder(sparse = False)
values_re = values.reshape(-1,1)
ohe = x_one_hot.fit_transform(values_re)
ohe_array = np.array(ohe)
ohe_column = pd.DataFrame(ohe_array)
ohe_column.columns = x_one_hot.get_feature_names()
data_drop = data.drop([‘soil’], axis = 1)
data_join_l = data_drop.join(ohe_column)
data_join = data_join_l

##split train and test set
group1_train_features = [‘b1’, ‘b2’, ‘b3’, ‘b4’, ‘2018max’,’2019max’, ‘ZZSXMC’, ‘dist’, ‘dvi2018x’, ‘ndwi2018x’, ‘rvi2018x’, ‘dem’,’hf_slope’, ‘DVI2019x’, ‘RVI2019x’]
label = [‘som’]
x = data[group1_train_features]
y = data[label]
x_train,x_test,y_train,y_test = train_test_split(x,y,train_size = 0.9, random_state = 0)

x = data[[‘b1’,’b2’,’b3’,’b4’]]
```

```
y = data['som']

x_train,x_test,y_train,y_test = train_test_split(x,y,test_size = 0.23, random_state = 0)


##xgb regression
model = xgb.XGBRegressor(random_state = 0, n_estimators = 14, made_depth = 2)
model.fit(x_train,y_train)

r_sq=model.score(x_test, y_test)
y_pre = model.predict(x_test)


def MSE(y, y_pre):
    return np.mean((y − y_pre) ** 2)
def RMSE(y, y_pre):
    return np.sqrt(MSE(y, y_pre))
def R2(y, y_pre):
    u = np.sum((y − y_pre) ** 2)
    v = np.sum((y − np.mean(y)) ** 2)
    return 1 − (u/v)


##print
print('r2',r_sq)
print('RMSE',RMSE(y_test,y_pre))
print('MSE',MSE(y_test,y_pre))
```

References

1. Mulla, D.J. Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosyst. Eng.* **2013**, *114*, 358–371. [CrossRef]
2. Lehmann, J.; Kleber, M. The contentious nature of soil organic matter. *Nature* **2015**, *528*, 60–68. [CrossRef] [PubMed]
3. Kellerman, A.M.; Kothawala, D.N.; Dittmar, T.; Tranvik, L.J. Persistence of dissolved organic matter in lakes related to its molecular characteristics. *Nat. Geosci.* **2015**, *8*, 454–457. [CrossRef]
4. Tuomisto, H.L.; Hodge, I.D.; Riordan, P.; Macdonald, D.W. Does organic farming reduce environmental impacts A meta-analysis of European research. *J. Environ. Manag.* **2012**, *112*, 309–320. [CrossRef]
5. Berhe, A.A.; Barnes, R.T.; Six, J.; Marin-Spiotta, E. Role of Soil Erosion in Biogeochemical Cycling of Essential Elements: Carbon, Nitrogen, and Phosphorus. *Annu. Rev. Earth Planet. Sci.* **2018**, *46*, 521–548. [CrossRef]
6. Kaiser, K.; Kalbitz, K. Cycling downwards—dissolved organic matter in soils. *Soil Biol. Biochem.* **2012**, *52*, 29–32. [CrossRef]
7. Melillo, J.M.; Frey, S.D.; DeAngelis, K.M.; Werner, W.J.; Bernard, M.J.; Bowles, F.P.; Pold, G.; Knorr, M.A.; Grandy, A.S. Long-term pattern and magnitude of soil carbon feedback to the climate system in a warming world. *Science* **2017**, *358*, 101–104. [CrossRef]
8. Caulfield, M.E.; Fonte, S.J.; Tittonell, P.; Vanek, S.J.; Sherwood, S.; Oyarzun, P.; Borja, R.M.; Dumble, S.; Groot, J.C.J. Inter-community and on-farm asymmetric organic matter allocation patterns drive soil fertility gradients in a rural Andean landscape. *Land Degrad. Dev.* **2020**, *31*, 2973–2985. [CrossRef]
9. Wang, Y.; Liu, G.; Zhao, Z. Spatial heterogeneity of soil fertility in coastal zones: A case study of the Yellow River Delta, China. *J. Soils Sediments* **2021**, *21*, 1826–1839. [CrossRef]
10. Jiang, Z.; Lian, F.; Wang, Z.; Xing, B. The role of biochars in sustainable crop production and soil resiliency. *J. Exp. Bot.* **2020**, *71*, 520–542. [CrossRef]
11. Ramesh, T.; Bolan, N.S.; Kirkham, M.B.; Wijesekara, H.; Kanchikerimath, M.; Rao, C.S.; Sandeep, S.; Rinklebe, J.; Ok, Y.S.; Choudhury, B.U.; et al. Soil organic carbon dynamics: Impact of land use changes and management practices: A review. In *Advances in Agronomy*; Sparks, D.L., Ed.; Academic Press: Cambridge, MA, USA, 2012; Volume 156, pp. 1–107.
12. Velasquez, E.; Lavelle, P. Soil macrofauna as an indicator for evaluating soil based ecosystem services in agricultural landscapes. *Acta Oecolog. Int. J. Ecol.* **2019**, *100*, 103446. [CrossRef]
13. Oldfield, E.E.; Wood, S.A.; Bradford, M.A. Direct effects of soil organic matter on productivity mirror those observed with organic amendments. *Plant Soil* **2018**, *423*, 363–373. [CrossRef]
14. Zhao, Y.N.; He, X.H.; Huang, X.C.; Zhang, Y.Q.; Shi, X.J. Increasing Soil Organic Matter Enhances Inherent Soil Productivity while Offsetting Fertilization Effect under a Rice Cropping System. *Sustainability* **2016**, *8*, 879. [CrossRef]

15. Yuan, X.; Chai, X.; Gao, R.; He, Y.; Jin, H.; Huang, Y. Temporal and spatial variability of soil organic matter in a county scale agricultural ecosystem. *N. Z. J. Agric. Res.* **2007**, *50*, 1157–1168. [CrossRef]
16. Hu, K.; Wang, S.; Li, H.; Huang, F.; Li, B. Spatial scaling effects on variability of soil organic matter and total nitrogen in suburban Beijing. *Geoderma* **2014**, *226*, 54–63. [CrossRef]
17. Huang, B.; Sun, W.; Zhao, Y.; Zhu, J.; Yang, R.; Zou, Z.; Ding, F.; Su, J. Temporal and spatial variability of soil organic matter and total nitrogen in an agricultural ecosystem as affected by farming practices. *Geoderma* **2007**, *139*, 336–345. [CrossRef]
18. van Beers, W.C.M.; Kleijnen, J.P.C. Kriging for interpolation in random simulation. *J. Oper. Res. Soc.* **2003**, *54*, 255–262. [CrossRef]
19. López-Granados, F.; Jurado-Expósito, M.; Peña-Barragán, J.M.; García-Torres, L. Using geostatistical and remote sensing approaches for mapping soil properties. *Eur. J. Agron.* **2005**, *23*, 279–289. [CrossRef]
20. Pouladi, N.; Møller, A.B.; Tabatabai, S.; Greve, M.H. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. *Geoderma* **2019**, *342*, 85–92. [CrossRef]
21. Henderson, T.L.; Szilagyi, A.; Baumgardner, M.F.; Chen, C.-C.T.; Landgrebe, D.A. Spectral Band Selection for Classification of Soil Organic Matter Content. *Soil Sci. Soc. Am. J.* **1989**, *53*, 1778–1784. [CrossRef]
22. Zhang, M.; Zhang, M.; Yang, H.; Jin, Y.; Zhang, X.; Liu, H. Mapping Regional Soil Organic Matter Based on Sentinel-2A and MODIS Imagery Using Machine Learning Algorithms and Google Earth Engine. *Remote. Sens.* **2021**, *13*, 2934. [CrossRef]
23. Santaga, F.S.; Agnelli, A.; Leccese, A.; Vizzari, M. Using Sentinel-2 for Simplifying Soil Sampling and Mapping: Two Case Studies in Umbria, Italy. *Remote Sens.* **2021**, *13*, 3379. [CrossRef]
24. Meng, X.; Bao, Y.; Ye, Q.; Liu, H.; Zhang, X.; Tang, H.; Zhang, X. Soil Organic Matter Prediction Model with Satellite Hyperspectral Image Based on Optimized Denoising Method. *Remote Sens.* **2021**, *13*, 2273. [CrossRef]
25. Nanni, M.R.; Demattê, J.A.; Rodrigues, M.; Santos, G.L.; Reis, A.S.; Oliveira, K.M.; Cezar, E.; Furlanetto, R.H.; Crusiol, L.G.; Sun, L. Mapping Particle Size and Soil Organic Matter in Tropical Soil Based on Hyperspectral Imaging and Non-Imaging Sensors. *Remote Sens.* **2021**, *13*, 1782. [CrossRef]
26. Gomez, C.; Rossel, R.A.V.; McBratney, A.B. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. *Geoderma* **2008**, *146*, 403–411. [CrossRef]
27. Li, X.-P.; Zhang, F.; Wang, X.-P. Study on Differential-Based Multispectral Modeling of Soil Organic Matter in Ebinur Lake Wetland. *Spectrosc. Spectr. Anal.* **2019**, *39*, 535–542. [CrossRef]
28. Zhai, M. Inversion of organic matter content in wetland soil based on Landsat 8 remote sensing image. *J. Vis. Commun. Image Represent.* **2019**, *64*, 102645. [CrossRef]
29. Baumgardner, M.F.; Kristof, S.; Johannsen, C.J.; Zachary, A. Effects of organic matter on the multispectral properties of soils. *Proc. Indiana Acad. Sci.* **1969**, *79*, 413–422.
30. Chen, Y.; Zhang, M.; Fan, D.; Fan, K.; Wang, X. Linear Regression between CIE-Lab Color Parameters and Organic Matter in Soils of Tea Plantations. *Eurasian Soil Sci.* **2018**, *51*, 199–203. [CrossRef]
31. Kodaira, M.; Shibusawa, S. Using a mobile real-time soil visible-near infrared sensor for high resolution soil property mapping. *Geoderma* **2013**, *199*, 64–79. [CrossRef]
32. Rodionov, A.; Welp, G.; Damerow, L.; Berg, T.; Amelung, W.; Paetzold, S. Towards on-the-go field assessment of soil organic carbon using Vis-NIR diffuse reflectance spectroscopy: Developing and testing a novel tractor-driven measuring chamber. *Soil Tillage Res.* **2015**, *145*, 93–102. [CrossRef]
33. Biney, J.K.M.; Boruvka, L.; Chapman Agyeman, P.; Nemecek, K.; Klement, A. Comparison of Field and Laboratory Wet Soil Spectra in the Vis-NIR Range for Soil Organic Carbon Prediction in the Absence of Laboratory Dry Measurements. *Remote Sens.* **2020**, *12*, 3082. [CrossRef]
34. Rasul, A.; Balzter, H.; Ibrahim, G.R.F.; Hameed, H.M.; Wheeler, J.; Adamu, B.; Ibrahim, S.a.; Najmaddin, P.M. Applying built-up and bare-soil indices from Landsat 8 to cities in dry climates. *Land* **2018**, *7*, 81. [CrossRef]
35. Xu, C.; Qu, J.J.; Hao, X.; Cosh, M.H.; Prueger, J.H.; Zhu, Z.; Gutenberg, L. Downscaling of surface soil moisture retrieval by combining MODIS/Landsat and in situ measurements. *Remote Sens.* **2018**, *10*, 210. [CrossRef]
36. Zhang, Y.; Guo, L.; Chen, Y.; Shi, T.; Luo, M.; Ju, Q.; Zhang, H.; Wang, S. Prediction of Soil Organic Carbon based on Landsat 8 Monthly NDVI Data for the Jianghan Plain in Hubei Province, China. *Remote Sens.* **2019**, *11*, 1683. [CrossRef]
37. Yu, H.; Liu, M.; Du, B.; Wang, Z.; Hu, L.; Zhang, B. Mapping Soil Salinity/Sodicity by using Landsat OLI Imagery and PLSR Algorithm over Semiarid West Jilin Province, China. *Sensors* **2018**, *18*, 1048. [CrossRef]
38. Fu, C.; Gan, S.; Yuan, X.; Xiong, H.; Tian, A. Impact of Fractional Calculus on Correlation Coefficient between Available Potassium and Spectrum Data in Ground Hyperspectral and Landsat 8 Image. *Mathematics* **2019**, *7*, 488. [CrossRef]
39. Seema; Ghosh, A.K.; Das, B.S.; Reddy, N. Application of VIS-NIR spectroscopy for estimation of soil organic carbon using different spectral preprocessing techniques and multivariate methods in the middle Indo-Gangetic plains of India. *Geoderma Reg.* **2020**, *23*, e00349. [CrossRef]
40. Dou, X.; Wang, X.; Liu, H.; Zhang, X.; Meng, L.; Pan, Y.; Yu, Z.; Cui, Y. Prediction of soil organic matter using multi-temporal satellite images in the Songnen Plain, China. *Geoderma* **2019**, *356*, 113896. [CrossRef]
41. Cao, X.; Li, X.; Ren, W.; Wu, Y.; Liu, J. Hyperspectral estimation of soil organic matter content using grey relational local regression model. *Grey Syst.Theory Appl.* **2020**, *11*, 707–722. [CrossRef]
42. Costa, E.M.; Tassinari, W.d.S.; Koenow Pinheiro, H.S.; Beutler, S.J.; Cunha dos Anjos, L.H. Mapping Soil Organic Carbon and Organic Matter Fractions by Geographically Weighted Regression. *J. Environ. Qual.* **2018**, *47*, 718–725. [CrossRef]

43. Takata, Y.; Funakawa, S.; Akshalov, K.; Ishida, N.; Kosaki, T. Spatial prediction of soil organic matter in northern Kazakhstan based on topographic and vegetation information. *Soil Sci. Plant Nutr.* **2007**, *53*, 289–299. [CrossRef]
44. Emadi, M.; Taghizadeh-Mehrjardi, R.; Cherati, A.; Danesh, M.; Mosavi, A.; Scholten, T. Predicting and Mapping of Soil Organic Carbon Using Machine Learning Algorithms in Northern Iran. *Remote Sens.* **2020**, *12*, 2234. [CrossRef]
45. Kobayashi, Y.; Yoshida, K. Quantitative structure?property relationships for the calculation of the soil adsorption coefficient using machine learning algorithms with calculated chemical properties from open-source software. *Environ. Res.* **2021**, *196*. [CrossRef]
46. Dong, Z.; Wang, N.; Liu, J.; Xie, J.; Han, J. Combination of machine learning and VIRS for predicting soil organic matter. *J. Soils Sediments* **2021**, *21*, 2578–2588. [CrossRef]
47. Wang, X.; Han, J.; Wang, X.; Yao, H.; Zhang, L. Estimating Soil Organic Matter Content Using Sentinel-2 Imagery by Machine Learning in Shanghai. *IEEE Access* **2021**, *9*, 78215–78225. [CrossRef]
48. Wang, Z.; Du, Z.; Li, X.; Bao, Z.; Zhao, N.; Yue, T. Incorporation of high accuracy surface modeling into machine learning to improve soil organic matter mapping. *Ecol. Indic.* **2021**, *129*, 107975. [CrossRef]
49. Yang, J.; Li, X.; Wu, B.; Wu, J.; Sun, B.; Yan, C.; Gao, Z. High Spatial Resolution Topsoil Organic Matter Content Mapping Across Desertified Land in Northern China. *Front. Environ. Sci.* **2021**, *9*, 668912. [CrossRef]
50. Hong, Y.; Liu, Y.; Chen, Y.; Liu, Y.; Yu, L.; Liu, Y.; Cheng, H. Application of fractional-order derivative in the quantitative estimation of soil organic matter content through visible and near-infrared spectroscopy. *Geoderma* **2019**, *337*, 758–769. [CrossRef]
51. Wang, Z.; Wang, G.; Zhang, G.; Wang, H.; Ren, T. Effects of land use types and environmental factors on spatial distribution of soil total nitrogen in a coalfield on the Loess Plateau, China. *Soil Tillage Res.* **2021**, *211*, 105027. [CrossRef]
52. Bokde, N.D.; Ali, Z.H.; Al-Hadidi, M.T.; Farooque, A.A.; Jamei, M.; Al Maliki, A.A.; Beyaztas, B.H.; Faris, H.; Yaseen, Z.M. Total Dissolved Salt Prediction Using Neurocomputing Models: Case Study of Gypsum Soil Within Iraq Region. *IEEE Access* **2021**, *9*, 53617–53635. [CrossRef]
53. Liu, L.; Ji, M.; Buchroithner, M. Combining Partial Least Squares and the Gradient-Boosting Method for Soil Property Retrieval Using Visible Near-Infrared Shortwave Infrared Spectra. *Remote Sens.* **2017**, *9*, 1299. [CrossRef]
54. Wang, Z.; Wang, G.; Zhang, Y.; Wang, R. Quantification of the effect of soil erosion factors on soil nutrients at a small watershed in the Loess Plateau, Northwest China. *J. Soils Sediments* **2020**, *20*, 745–755. [CrossRef]
55. Ahirwal, J.; Nath, A.; Brahma, B.; Deb, S.; Sahoo, U.K.; Nath, A.J. Patterns and driving factors of biomass carbon and soil organic carbon stock in the Indian Himalayan region. *Sci. Total Environ.* **2021**, *770*, 145292. [CrossRef] [PubMed]
56. Jiang, G.; Grafton, M.; Pearson, D.; Bretherton, M.; Holmes, A. Predicting spatiotemporal yield variability to aid arable precision agriculture in New Zealand: A case study of maize-grain crop production in the Waikato region. *N. Z. J. Crop. Hortic. Sci.* **2021**, *49*, 41–62. [CrossRef]
57. Li, M.; Xi, X.; Xiao, G.; Cheng, H.; Yang, Z.; Zhou, G.; Ye, J.; Li, Z. National multi-purpose regional geochemical survey in China. *J. Geochem. Explor.* **2014**, *139*, 21–30. [CrossRef]
58. Tian, F.; Wang, Y.; Fensholt, R.; Wang, K.; Zhang, L.; Huang, Y. Mapping and Evaluation of NDVI Trends from Synthetic Time Series Obtained by Blending Landsat and MODIS Data around a Coalfield on the Loess Plateau. *Remote Sens.* **2013**, *5*, 4255–4279. [CrossRef]
59. Ma, Y.; Liu, H.; Jiang, B.; Meng, L.; Guan, H.; Xu, M.; Cui, Y.; Kong, F.; Yin, Y.; Wang, M. An Innovative Approach for Improving the Accuracy of Digital Elevation Models for Cultivated Land. *Remote Sens.* **2020**, *12*, 3401. [CrossRef]
60. Busch, R.; Hardt, J.; Nir, N.; Schuett, B. Modeling Gully Erosion Susceptibility to Evaluate Human Impact on a Local Landscape System in Tigray, Ethiopia. *Remote Sens.* **2021**, *13*, 2009. [CrossRef]
61. Zhao, C.; Zhou, Y.; Jiang, J.H.; Xiao, P.N.; Wu, H. Spatial characteristics of cultivated land quality accounting for ecological environmental condition: A case study in hilly area of northern Hubei province, China. *Sci. Total Environ.* **2021**, *774*, 145765. [CrossRef]
62. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
63. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
64. Ma, X.; Sha, J.; Wang, D.; Yu, Y.; Yang, Q.; Niu, X. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electron. Commer. Res. Appl.* **2018**, *31*, 24–39. [CrossRef]
65. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Volume 10, pp. 785–794.
66. Giannakas, F.; Troussas, C.; Krouska, A.; Sgouropoulou, C.; Voyiatzis, I. XGBoost and Deep Neural Network Comparison: The Case of Teams' Performance. In *International Conference on Intelligent Tutoring Systems*; Springer: Cham, Switzerland, 2021; pp. 343–349.
67. Zamani Joharestani, M.; Cao, C.; Ni, X.; Bashir, B.; Talebiesfandarani, S. PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. *Atmosphere* **2019**, *10*, 373. [CrossRef]
68. Zhou, L.; Luo, T.; Du, M.; Chen, Q.; Liu, Y.; Zhu, Y.; He, C.; Wang, S.; Yang, K. Machine Learning Comparison and Parameter Setting Methods for the Detection of Dump Sites for Construction and Demolition Waste Using the Google Earth Engine. *Remote Sens.* **2021**, *13*, 787. [CrossRef]
69. Zhang, Y.; Ma, J.; Liang, S.; Li, X.; Li, M. An Evaluation of Eight Machine Learning Regression Algorithms for Forest Aboveground Biomass Estimation from Multiple Satellite Data Products. *Remote Sens.* **2020**, *12*, 4015. [CrossRef]

70. Ramezan, C.; Warner, T.; Maxwell, A. Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. *Remote Sens.* **2019**, *11*, 185. [CrossRef]
71. Wang, Z.; Hu, M.; Zhai, G. Application of Deep Learning Architectures for Accurate and Rapid Detection of Internal Mechanical Damage of Blueberry Using Hyperspectral Transmittance Data. *Sensors* **2018**, *18*, 1126. [CrossRef]
72. Fan, M.; Lal, R.; Zhang, H.; Margenot, A.J.; Wu, J.; Wu, P.; Zhang, L.; Yao, J.; Chen, F.; Gao, C. Variability and determinants of soil organic matter under different land uses and soil types in eastern China. *Soil Tillage Res.* **2020**, *198*, 104544. [CrossRef]
73. Zhisheng, A.; Tunghseng, L.; Yanchou, L.; Porter, S.C.; Kukla, G.; Xihao, W.; Yingming, H. The long-term paleomonsoon variation recorded by the loess-paleosol sequence In central China. *Quat. Int.* **1990**, *7*, 91–95. [CrossRef]
74. Villamil-Cubillos, L.F.; Leon-Medina, J.X.; Anaya, M.; Tibaduiza, D.A. Evaluation of Feature Selection Techniques in a Multifrequency Large Amplitude Pulse Voltammetric Electronic Tongue. *Eng. Proc.* **2020**, *2*, 62. [CrossRef]
75. Zhang, W.; Wu, C.; Zhong, H.; Li, Y.; Wang, L. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geosci. Front.* **2021**, *12*, 469–477. [CrossRef]
76. Guo, N.; Shi, X.; Zhao, Y.; Xu, S.; Wang, M.; Zhang, G.; Wu, J.; Huang, B.; Kong, C. Environmental and anthropogenic factors driving changes in paddy soil organic matter: A case study in the Middle and Lower Yangtze River Plain of China. *Pedosphere* **2017**, *27*, 926–937. [CrossRef]
77. Duan, L.; Li, Z.; Xie, H.; Li, Z.; Zhang, L.; Zhou, Q. Large-scale spatial variability of eight soil chemical properties within paddy fields. *Catena* **2020**, *188*, 104350. [CrossRef]
78. Wang, H.; Zhang, X.; Wu, W.; Liu, H. Prediction of Soil Organic Carbon under Different Land Use Types Using Sentinel-1/-2 Data in a Small Watershed. *Remote Sens.* **2021**, *13*, 1229. [CrossRef]
79. Li, Z.Q.; Zhang, X.; Xu, J.; Cao, K.; Wang, J.H.; Xu, C.X.; Cao, W.D. Green manure incorporation with reductions in chemical fertilizer inputs improves rice yield and soil organic matter accumulation. *J. Soils Sediments* **2020**, *20*, 2784–2793. [CrossRef]
80. Du, Z.; Gao, B.; Ou, C.; Du, Z.; Yang, J.; Batsaikhan, B.; Dorjgotov, B.; Yun, W.; Zhu, D. A Quantitative Analysis of Factors Influencing Organic Matter Concentration in the Topsoil of Black Soil in Northeast China Based on Spatial Heterogeneous Patterns. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 348. [CrossRef]
81. Sheng, Y.; Liu, W.; Xu, H.; Gao, X. The Spatial Distribution Characteristics of the Cultivated Land Quality in the Diluvial Fan Terrain of the Arid Region: A Case Study of Jimsar County, Xinjiang, China. *Land* **2021**, *10*, 896. [CrossRef]
82. Sahabiev, I.; Smirnova, E.; Giniyatullin, K. Spatial Prediction of Agrochemical Properties on the Scale of a Single Field Using Machine Learning Methods Based on Remote Sensing Data. *Agronomy* **2021**, *11*, 2266. [CrossRef]
83. Wang, X.; Shi, W.; Sun, X.; Wang, M. Comprehensive benefits evaluation and its spatial simulation for well-facilitated farmland projects in the Huang-Huai-Hai Region of China. *Land Degrad. Dev.* **2020**, *31*, 1837–1850. [CrossRef]
84. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015; pp. 1–9.
85. Ciresan, D.; Giusti, A.; Gambardella, L.; Schmidhuber, J. Deep neural networks segment neuronal membranes in electron microscopy images. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 2843–2851.
86. Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; Müller, K.-R. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit.* **2017**, *65*, 211–222. [CrossRef]