Identification of Soil Texture Classes Under Vegetation Cover Based on Sentinel-2 Data With SVM and SHAP Techniques

Yanan Zhou, Wei Wu, Huan Wang, Xin Zhang, Chao Yang, and Hongbin Liu

***Abstract*—Understanding the spatial variability of soil texture classes is essential for agricultural management and environment sustainability. Sentinel-2 data offer valuable vegetation information as proxies for soil properties inference. However, the applications of them in soil texture classification are still limited. This study investigated the usefulness of Sentinel-2 data for predicting soil texture class using an interpretable machine learning (ML) strategy. Specifically, multitemporal Sentinel-2 images were used to get exhaustive vegetation cover information. Basic digital elevation map (DEM) derivatives and stratum were extracted. Three support vector machines with different input parameters (purely DEM derivatives and stratum, purely Sentinel-2, and Sentinel-2 plus DEM derivatives and stratum) were developed. Moreover, in order to improve the transparency in black box ML models, the novel SHapley Additive exPlanations (SHAP) method was applied to interpret the outputs and analyze the importance of individual variables. Results showed that the model with all variables provided desirable performance with overall accuracy of 0.8435, F1-score of 0.835, kappa statistic of 0.7642, precision of 0.8388, recall of 0.8355, and area under the curve of 0.9451. The model with purely Sentinel-2 data performed much better than that with solely DEM derivatives and stratum. The contributions of Sentinel-2 data to explain soil texture class variability were about 17%, 41%, and 28% for sandy, loamy and clayey soils, respectively. The SHAP method visualized the decision process of ML and indicated that elevation, stratum, and red-edge factors were critical variables for predicting soil texture classes. This study offered much-needed insights into the applications of Sentinel-2 data in digital soil mapping and ML-assisted tasks.**

***Index Terms*—Machine learning (ML), sentinel-2, shapley additive explanations (SHAP), soil texture classes.**

I. Introduction

Soil texture controls soil water holding capacity, permeability, solute movement and aeration [1], and then drives soil physical, chemical, biological, and hydrological processes that are closely related to plant growth and soil erosion [2], [3]. Soil texture also determines the suitability of soil for agricultural production [1]. Thus, understanding explicitly spatial variability of soil texture is necessary for ecosystem services, environment sustainability, and smart agricultural management [4], [5].

Compared with the time-consuming and expensive field surveys, digital soil mapping (DSM) has been developed for predicting soil properties due to its reliability and high efficiency [6]–[14]. Remote sensing images are attractive and essential data in DSM because of high revisit rate, resolution and availability at a range of scales, timeliness, low cost, and convenience [6]–[14]. In terms of soil texture, most scholars use remote sensing to obtain texture information based on spectroscopy about the reflection of bare soil pixels [7], [11]–[14]. These researches rely heavily on the availability of bare soils and are confined to the cropland under fallow or seedbed conditions [7], [11], arid and semi-arid regions with sparse vegetation [12], or areas with bare soils identified by time series multispectral images [13], [14]. Given that satellites images (e.g., Landsat, MODIS) can also characterize vegetation properties as proxies for soil attributes inference [9], [15], some works have started to reach soil texture using remote-sensed vegetation indices with other auxiliary variables (e.g., topography, stratum) in densely vegetated areas (e.g., forest system, tropical hillslope environment) [9], [16]–[18]. These studies demonstrate that the addition of satellite covariates representing vegetation properties could improve soil texture prediction accuracy [16] and multitemporal optical images recording abundant crop growth information are useful for identifying soil texture classes [9]. In the subtropical monsoon climate zone, intense cultivation is universal and hence it is difficult to capture signals of exposed soils. More exhaustive vegetation information is needed to map soil texture spatial variation. Sentinel-2, the new earth observation satellites open freely to users, are promising in crop identification, vegetation types classification and biomass mapping due to their higher spatiotemporal resolution (10 or 20 m, 5-day revisit cycle) and richer spectral channels (thirteen bands) [19]–[21]. Therefore, we hypothesized that Sentinel-2 data could offer detailed vegetation cover information to express soil texture variability.

Based on the DSM framework [22], various statistical techniques have been used to predict soil properties including multiple linear regression [23], geostatistical methods [24], and machine learning (ML) modelings [25]–[28]. In most cases, ML algorithms show better prediction performance because they

Manuscript received October 25, 2021; revised March 17, 2022; accepted March 26, 2022. Date of publication April 1, 2022; date of current version May 19, 2022. *(Corresponding author: Hongbin Liu.)*

Yanan Zhou, Huan Wang, Xin Zhang, and Hongbin Liu are with the College of Resources and Environment, Southwest University, Chongqing 400716, China (e-mail: zyn1999@email.swu.edu.cn; wh22790@163.com; zhang_xin1211@163.com; liuhongbinswu@163.com).

Wei Wu is with the College of Computer and Information Science, Southwest University, Chongqing 400716, China (e-mail: wuwei_star@163.com).

Chao Yang is with the Chongqing Tobacco Science Institute, Southwest University, Chongqing 400716, China (e-mail: 358235772@qq.com).

Digital Object Identifier 10.1109/JSTARS.2022.3164140

can learn nonlinear interactions iteratively from data without potential loss of related information [29], such as random forest (RF) [25], classification and regression tree [26] and support vector machine (SVM) [27], [28]. SVM is famous for robust generalization performance in ML algorithms [27], [28]. It seeks the best compromise between the complexity of the model and fitting precision based on the structure risk minimization principle and has a well-established theoretical fundamental. Therefore, SVM is often favored in DSM [27], [28].

Although ML algorithms could handle the nonlinear relationship between target and predictor variables, the interpretability of them remains a challenge due to inherent "black box" property. The best explanation for a simple model is the model formula itself, but it is not feasible for ML algorithms. Thus, the goal of current interpretation strategies for ML algorithm is not to explain the logical concept underlying the black box, but to give the reasonable reasons for the estimated value of a special instance [30]. Traditional variables contribution methods (e.g., permutation, gini) tend to evaluate the global importance of variables to understand model outputs, which are not individualized for each prediction, causing local inconsistency [31]. Recently, a new SHapley Additive exPlanations (SHAP) approach was proposed to estimate the contribution of an individual variable by comparing the performance of the model with and without the variable [32]. Interestingly, the contribution of variables in each instance is also calculated. Therefore, both local and global importance of input variables are generated for the response [33]. As a tool for providing better explanations for model outputs, SHAP approach has broad application prospects for predicting natural and social phenomena [34]–[37]. To our knowledge, there are no reports in the literature about the application of SHAP method into DSM framework.

Therefore, the main objectives of the current study were to 1) explore the potential of Sentinel-2 data for identifying soil texture classes by using retrieved vegetation properties under vegetation cover conditions and 2) analyze the critical factors affecting soil texture class variation with SHAP method. Specifically, three SVMs with different input parameters [purely digital elevation map (DEM) derivatives and stratum, purely Sentinel-2, and Sentinel-2 plus DEM derivatives and stratum] were evaluated to investigate the contribution of Sentinel-2 data to explain soil texture class spatial distribution, in consideration of the impacts of other environmental factors (e.g., terrain, geology) on soil texture class variability.

II. Materials and Methods

A. Study Area

The study area is located in southwestern China (see Fig. 1). It is a small basin (6790 ha) in the Three Gorges region of the Yangtze River. The climate is subtropical monsoon humid. The study site is often covered by clouds. The average annual temperature is 15 °C ranging from 6 °C (January) and 36 °C (August). The mean precipitation is 1224 mm with 70% occurring in May to September. The topography is mainly mountains, with elevation varying between 146 and 1625 m and slope ranging between 0° and 62°. Main land use types are forest and dry farmland accounting for 48.6% and 18.4% of the total area, respectively. Others are paddy field, water, grassland, and building. The dry farmland is under a rotational tillage farming system, with winter rapeseed (*Brassica napus L.*), corn (*Zea mays L.*), and sweet potato (*Ipomoea batatas L.*) for a long time. Corn is sown in April and reaped in July. Sweet potato is seeded in June and harvested in October. Winter rapeseed is planted in October and matures in the following April. Agricultural techniques and varieties are the same for each crop.

The soil parent materials over the study area come from two strata, namely, the Daye Formation deposited in the late Triassic and the Xujiahe Formation in the early Triassic [38]. The Xujiahe Formation is composed of numerous types of rock, including siltstone, fine sandstone, mudstone, etc. The Daye Formation is dominated by marls and thinly-bedded limestone. According to FAO soil classification, soils developed from the two geological units are classified as Regosols (the Xujiahe Formation) and Entisol (the Daye Formation), respectively [39]. Most soils are neutral with pH ranging from 5.4 to 8.6 and organic matter varying between 6.3 and 42 g/kg, respectively.

B. Soil Data

A total of 943 soil samples (0–20 cm) were collected from dry farmland in September 2012. According to the requirements of NY/T 1634–2008 Technical Regulations for National Cultivated Land Fertility Survey and Quality Evaluation, sampling locations were carefully selected considering the natural condition over the study area, such as geological substrate, topographic characteristics, and soil types. At each sampling point, 10 subsamples were randomly collected within a radius of 10 m around the point and then thoroughly mixed as the final sample. The soil texture classes were estimated by experienced technicians using the "Fingerprobe" method, which is regarded as an appropriate alternative to the determination of soil texture in the laboratory [40]–[42].

Soil texture could be divided into very fine classes (twelve) according to the United States Department of Agriculture (USDA). However, soil texture is always classified into three or four groups in practice [43]–[45]. In the current study site, three general classes of soil texture, namely, sandy, loamy, and clayey textures were identified [44].

In order to evaluate the accuracy about the estimated textural classes by Fingerprobe method in field, the particle sizes of 43 samples were analyzed using "Robinson" pipette method. Then the corresponding textural class was identified by soil texture calculator[1] (see Fig. 15). The result showed that the overall accuracy (OA) and kappa statistic of general classes using Fingerprobe method were 76.7% and 0.604, respectively (more details in Table IV). Thus, the data were suitable for further study.

The number of samples was 54, 502, and 387 for sandy, loamy, and clayey textural classes, respectively. For the Xujiahe Formation, the number of samples was 4, 258, and 83 for sandy, loamy, and clayey textural classes, respectively. For the Daye Formation, the number of samples was 50, 244, and 304 for sandy, loamy, and clayey textural classes, respectively.

[1] [Online]. Available: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs142p2_054167

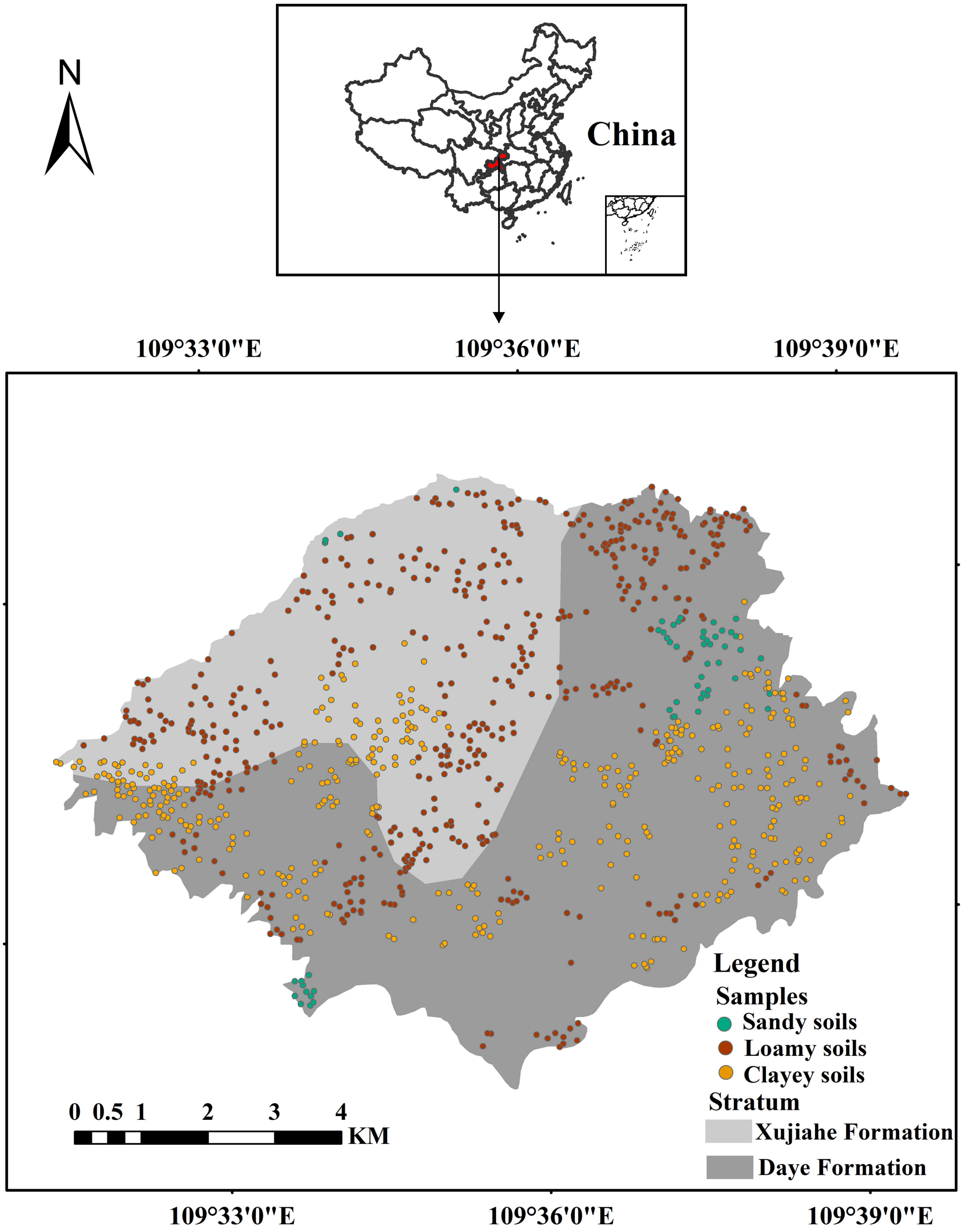

Fig. 1. Location, stratum, and sampling sites of the study area.

C. Predictor

The predictor variables used in this study included stratum [26], [43], DEM derivatives, and spectral indictors originated from Sentinel-2 imageries. The stratum was a categorical variable, which was processed by one-hot-encoded, and others were continuous variables. The variable values were extracted into the soil samples using the ArcGIS 10.6 software. Each variable was processed by min-max normalization according to its respective ranges.

1) Geological and Topographical Factors: The Shuttle Radar Topography Mission DEM with 30 m resolution[2] was used to provide consistent, high-quality elevation data [46]. This dataset was coregistered to the Sentinel-2 imageries and resampled to 10 m spatial resolution using nearest method based on ArcGIS 10.6 software. Then, the four basic DEM derivatives including elevation, slope, aspect, and topographic wetness index (TWI) were extracted by using SAGA GIS 7.2.0 software (see Fig. 2). Stratum was obtained from a geological map with a scale of 1:50000 (see Fig. 1).

2) Remote Sensing Data: There are few suitable images due to the frequent cloud over our study area. In order to minimize the noise of surface moisture resulting from rainfall [47], images that are away from rainfall dates are considered. Finally, only five cloud-free images covering spring (April 16th, 2019), summer (August 29th and September 20th, 2018),

[2] [Online]. Available: https://earthexplorer.usgs.gov/

TABLE I
SPECIFIC LIST OF REMOTE SENSING VARIABLES

AB	Remote sensing variable	Remarks	Reference
B5	Band 5	Red-edge band	[25,48,50]
B6	Band 6	Red-edge band	[25,48,50]
B7	Band 7	Red-edge band	[25,48,50]
IRECI	Inverted Red-Edge Chlorophyll Index^{1}	(B7-B4$^{\#}$)/(B5/B6)	[51]
REIP	Red-Edge Inflection Point^{1}	700+40*((B4+B7)/2-B5)/(B6-B5)	[51,52]
S2REP	Sentinel-2 Red-Edge Position Index^{1}	700+35*((B4+B7)/2-B5)/(B6-B5)	[51,52]
MCARI	Modified Chlorophyll Absorption in Reflectance Index^{1}	[(B5-B4)-0.2*(B5-B3$^{\#}$)*(B5/B4)]	[51,53]
NDI45	Normalized Difference Index^{1}	(B5-B4)/(B5+B4)	[51,54]
CI	Color Index^{2}	(B4-B3)/(B4+B3)	[9,48]
BI	Brightness Index^{2}	Sqrt((B4+B3)/2)	[9,48]
NDVI	Normalized Difference Vegetation Index^{1}	(B8$^{\#}$-B4)/(B8+B4)	[9,48,49]
NDWI	Normalized Difference Water Index^{1}	(B8-B12)/(B8+B12)	[1]
SAVI	Soil Adjusted Vegetation Index^{1}	(1+0.5)*(B8-B4)/(B8+B12+0.5)	[11,16]
B8A	Band 8A	Near-InfraRed	[55]
B11	Band 11	Short Wave InfraRed	[43,55,56]
B12	Band 12	Short Wave InfraRed	[43,55,56]

Note: Superscript number 1 denotes vegetation radiometric indices and 2 denotes soil radiometric indices. Superscript # denotes abbreviated bands not explained in the table: B3 is band 3 (visible light band, green band), B4 is band 4 (visible light band, red light band) and B8 is band 8 (near-infrared band).

TABLE II
NUMBER OF SAMPLES FOR EACH SOIL TEXTURE CLASS BEFORE AND AFTER SMOTE

Dataset	Sandy soils		Loamy soils		Clayey soils		Total
	Number	%	Number	%	Number	%	
Original	54	5.73	502	53.23	387	41.04	943
SMOTE	502	33.33	502	33.33	502	33.33	1506

TABLE III
EXAMPLE OF CONFUSION MATRIX

	Actual positive class	Actual negative class
Predict positive class	True positive (TP)	False positive (FP)
Predict negative class	False negative (FN)	True negative (TN)

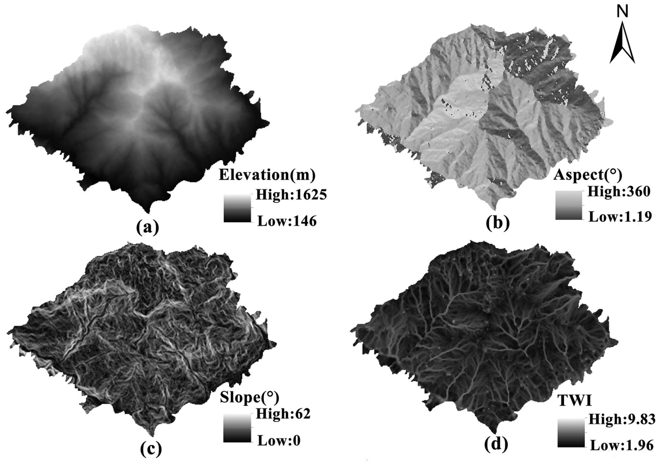

Fig. 2. Maps of (a) elevation, (b) aspect, (c) slope, and (d) TWI.

autumn (October 18th, 2018), and winter (December 12th, 2019) were downloaded from European Space Agency (ESA[3]). Each image covered individually the study area and included thirteen spectral bands in the VNIR–SWIR spectral domain with 10–60 m spatial resolutions. Atmospheric and topographic corrections were performed on these images through the Sen2cor plug-in unit provided by ESA. For each image, six spectral bands with 20 m spatial resolution were resampled to 10 m by bilinear interpolation method and three spectral bands with 60 m resolution were eliminated. These Sentinel-2 images were then clipped to cover the study area. All operations were done with Sentinel Application Platform (SNAP) 7.0.0 software.

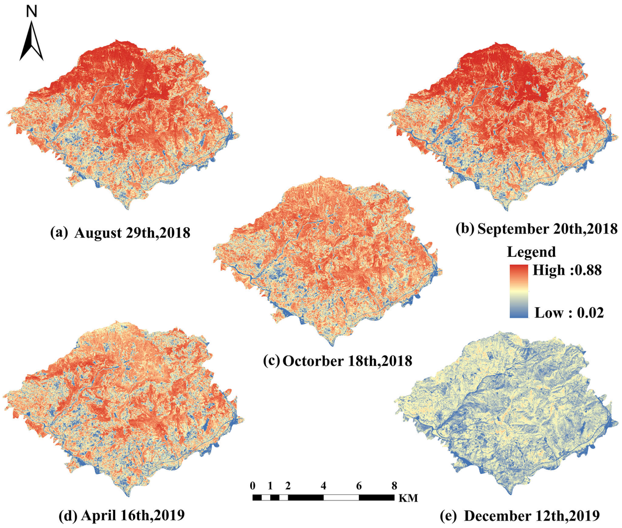

Fig. 3. Images of Normalized Difference Vegetation Index (NDVI) in different time phases over the study area.

Fig. 3 shows NDVI values over the study area in five different time phases. Based on these images, bare soil pixels were retrieved by the threshold of NDVI < 0.2 [48]. No pixels always had NDVI < 0.2 in these time phases. For each image, bands 5, 6, 7, 8A, 11, 12, and ten spectral indices were used to obtain exhaustive vegetation cover information to infer soil texture in the current study (see Table I) [9], [25], [48]–[55]. Finally, a total of 80 indices were derived from the five Sentinel-2 images.

D. Methods

The flowchart shown in Fig. 4 summarizes the methodology used in this work. It consists of three main steps:

[3] [Online]. Available: https://scihub.copernicus.eu/dhus/#/home

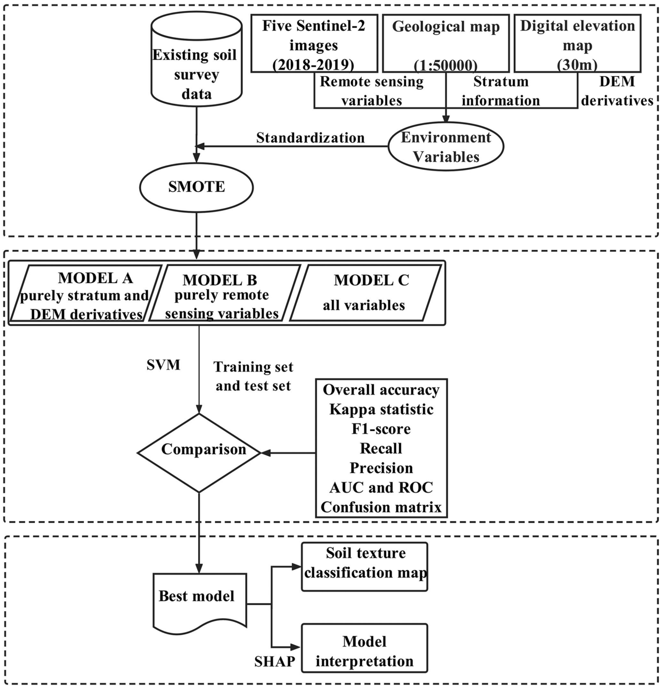

Fig. 4. Flowchart for methodology.

1) data preprocessing, which includes the balance of data set based on synthetic minority oversampling technique (SMOTE) and variables election using recursive feature elimination with cross validation (RFECV);
2) model construction and evaluation, where three SVMs with different inputs are compared to evaluate the potential of adding Sentinel-2 data for soil texture classification;
3) model interpretation and application, where soil texture classes are identified based on the best model and SHAP method is applied to interpret the results.

1) SMOTE: ML methods usually work well on balanced datasets. However, models could give a very poor performance on an imbalanced dataset [57]–[60]. In order to solve this problem, many techniques have been proposed. Among them, SMOTE has been successfully used in different fields [57]–[60]. SMOTE performs oversampling by creating synthetic examples in the variable space. Specifically, a random example from the minority class is first chosen. Then *k* nearest neighbors of that example are found. A linear function is generated between that example and each nearest neighbor, and a new synthetic sample is created by this function (see Fig. 5).

In this case, the imbalanced data were converted into a new balance dataset using SMOTE_NC in Imblearn 0.7.0 package of Python 3.6.10 [61]. The original and oversampling datasets were shown in Table II.

2) Variable Selection: The RFECV was used to conduct variable selection for simplifying statistical problems and eliminating redundant information. RFECV performs a recursive variables elimination process by identifying and removing variables with low importance. Meanwhile, it applies a cross-validation method to find an optimal subset with minimum generalization errors [37]. RF algorithm (parameters mtry = default (square root of the number of input features) and ntree = 500) was used to calculate the importance of each variable based on Mean Decrease Impurity [62], because it was insensitive to non-informative predictors [62], [63]. The number of variables was determined automatically based on the highest OA. Finally, a total of 47 variables were retained, including all geological and topographical variables and 42 remote sensing variables (see Fig. 6 and Table V). These variables were used in the following analyses.

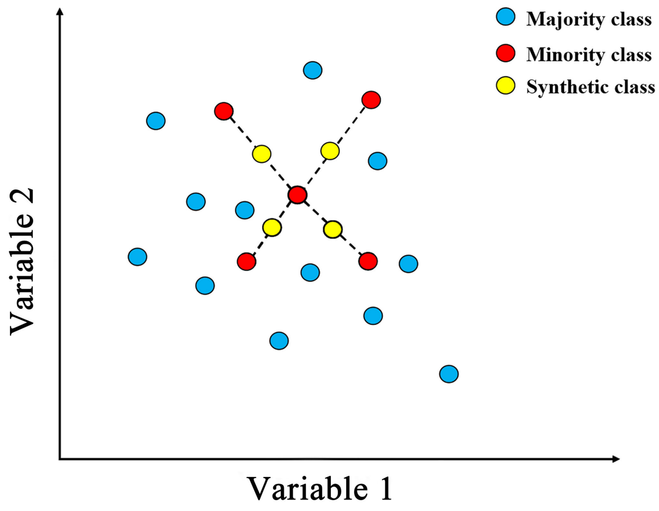

Fig. 5. Illustrative overview of the SMOTE.

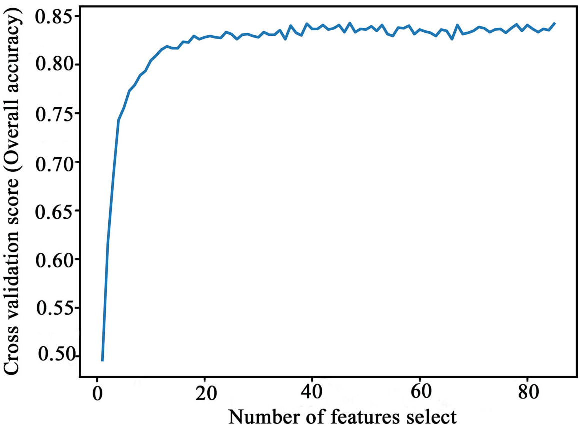

Fig. 6. Optimal variables subset selection using RFECV-random forest.

3) Support Vector Machine: SVM is a popular supervised classification and regression learning technology. It is by nature a binary classifier, but it can be extended for multiclass classification using the “one-against-all” (ovr) or “one-against-one" (ovo) strategies. The kernel functions extend SVM to a nonlinear model. There are four typical kernel functions including polynomial, sigmoid, linear, and radial basis function (RBF). SVM with RBF was used in this case, since it has been widely used in DSM [27], [28]. The SVR model of scikit-learn 0.23.0 implemented with Python 3.6.10 was used in this study. The kernel function was “RBF” and the classification strategy was “ovo.”

4) Model Evaluation: Three SVM models (MODEL A, B, and C, hereafter) with different input combinations were created in the current work. MODEL A solely included DEM derivatives and stratum, MODEL B purely contained Sentinel-2 images, and MODEL C had all variables.

We selected randomly 75% of the data for training and the remainders for test. Ten-fold cross-validation was applied to optimize model parameters based on the training sets. Specifically, the training set was divided into ten pairs of mutually exclusive subsets, of which nine pairs were used as training sets and the remaining were verification sets. The experiment was performed with the subsets in turn, and ten verification results were averaged. Two parameters including penalty (cost) and kernel width (gamma) are needed for the RBF. The appropriate parameters (cost and gamma) were selected according to the mean OA. GridSearch-CV module in scikit-learn 0.23.0 of Python 3.6.10 was used to search optimal parameters of SVM in the parameter space (gamma $\in$ (0,10], cost $\in$[1, 100]). Values of cost and gamma were 20 and 5, 9 and 3, 10.65 and 0.75 for MODEL A, B, and C, respectively.

The model performance was evaluated based on the training and independent test sets. The evaluation indexes included OA, kappa statistic, precision, recall rate, and F1-score. The calculation formulas are as follows [60]:

$$\mathrm{OA} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} \tag{1}$$

$$P_e = \frac{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})+(\mathrm{FP}+\mathrm{TN})(\mathrm{FN}+\mathrm{TN})}{\mathrm{N}^2} \tag{2}$$

$$\mathrm{Kappa} = 1 - \frac{1-\mathrm{OA}}{1-P_e} \tag{3}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{4}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{5}$$

$$F1-\mathrm{score} = \frac{\mathrm{TP}\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{6}$$

where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively (see Table III). *N* is the sample number. Models with higher values of OA, Kappa, Precision, Recall, and F1-score perform better. Model performance indicated by Kappa is: <0, poor; 0.00–0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1.00, almost perfect [64].

These indicators were calculated by the confusion matrix. An example of a confusion matrix was shown in Table III.

In addition, the receiver operating characteristic curve (ROC) was drawn and the area under the curve (AUC) was calculated. ROC curve shows the relationship between true positive rate (TPR) and false positive rate (FPR). TPR and FPR are as follows [25]:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{7}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}\,. \tag{8}$$

An AUC value of 1 suggests a perfect model, while an AUC value of 0 indicates a noninformative model [46]. According to AUC, the model performs excellent (0.9–1), very good (0.8–0.9), good (0.7–0.8), average (0.6–0.7), and poor (0.5–0.6) [65].

5) Model Interpretation: Complex ML technique offers a superior capability to account for nonlinear relationships between predictors and target and interaction effects between predictors, but it is limited by the lack of interpretability and “black box” properties. The SHAP proposed by Lundberg and Lee in 2017 [32] furnishes a unique solution for interpreting model output. SHAP method is based on game theory and local explanations, satisfying the following properties: local accuracy, missingness, and consistence [66]–[68]. This method calculates the shapley value for variable *i* to estimate its contribution to model output (v(N)) using the formula [69]:

$$\emptyset_i = \sum_{S\subseteq M\backslash\{i\}} \frac{|m!|\,(m-|s|-1)!}{m!}\left[(v\,(S\cup\{i\})-v\,(S)\right] \tag{9}$$

where $\emptyset_i$ denotes the shapley value of variable *i*. *S* represents the variable subset that does not contain variable *i*. *M* is the set of all input variables and m is the number of variables.

A linear function of binary variables *g* is defined to replace the original model *f*:

$$\mathrm{f\,(x)} = \mathrm{g\,(x')} = \emptyset_0 + \sum_{i=1}^{m} \emptyset_i \tag{10}$$

where $\emptyset_0$ is the constant when all inputs are missing. For each sample, the model output is decomposed into the sum of a constant and the shapley values (contribution) of all variables. Then, both local and global importance of input variables are obtained. SHAP library v0.37.0[4] embedded in Python 3.6.10 was used in this study [34], [36].

III. Results

A. Model Performance

Fig. 7 shows the model performance evaluated using training and test datasets before and after SMOTE. For each model, similar values of statistical accuracy indicators were produced by both training and test datasets before or after SMOTE. This confirmed the robust generalization ability of SVM. Meanwhile, model performances displayed accordant trends for the datasets before and after SMOTE. That is, MODEL A with solely DEM derivatives and stratum performed worst, MODEL B with purely Sentinel-2 data gave much higher classification accuracy than MODEL A, and MODEL C with all parameters performed best. Also, the relative improvement of MODEL C over MODEL A was more prominent than that of MODEL C over MODEL B. These demonstrated that the addition of Sentinel-2 data can bring significant amelioration in prediction accuracy than the addition of topography and stratum parameters. Moreover, the comparison of various accuracy indicators for the same model before and after SMOTE shows the benefits of using SMOTE, which was coincident with some literatures [57], [58], [60]. Therefore, the following analyses were based on the results after SMOTE.

Fig. 8 shows the AUC scores of the three models calculated with the test dataset. According to the AUC scores, MODEL C performed excellently, MODEL A and MODEL B performed very good. The higher AUC scores (exceeded 0.8) of the three models pointed out that these different types of environment

[4] [Online]. Available: https://github.com/slundberg/shap

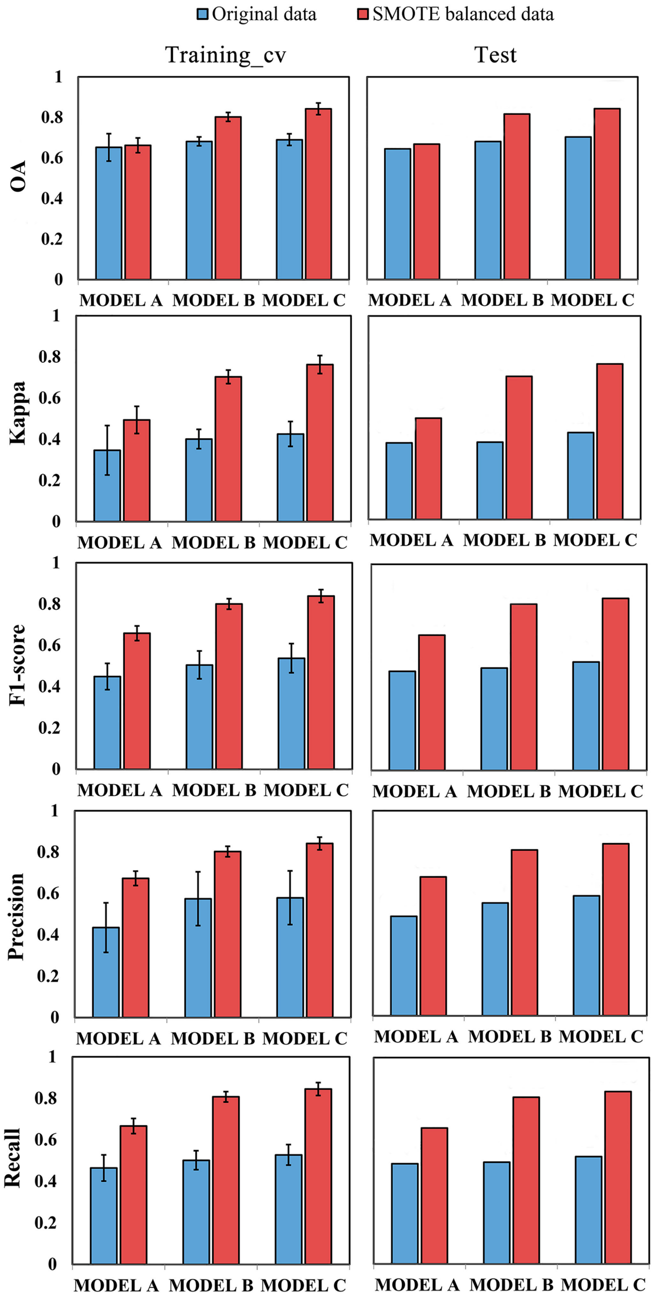

Fig. 7. Various assessment scores for Models A, B, and C on training and test dataset before and after SMOTE (Training_cv denotes ten-fold cross validation of training dataset. The left side shows the cross-validation results of the training set and the right side shows the results of the test set. MODEL A: solely DEM derivatives, stratum. MODEL B: purely Sentinel-2 data. MODEL C: Sentinel-2 data plus DEM derivatives and stratum. The same below).

variables (Sentinel-2 data, DEM derivatives, and stratum) were suitable for identifying soil texture classes.

The confusion matrices further concretely reveal the classification results of the three models (see Fig. 9). MODEL A with solely DEM derivatives and stratum could identify about 85% sandy, 50% loamy, and 63% clayey textural soils. MODEL B with purely Sentinel-2 data could discern about 99% sandy, 66% loamy, and 78% clayey textural soils. MODEL C with DEM derivatives, stratum, and Sentinel-2 data could distinguish about 99% sandy, 70% loamy, and 81% clayey textural soils.

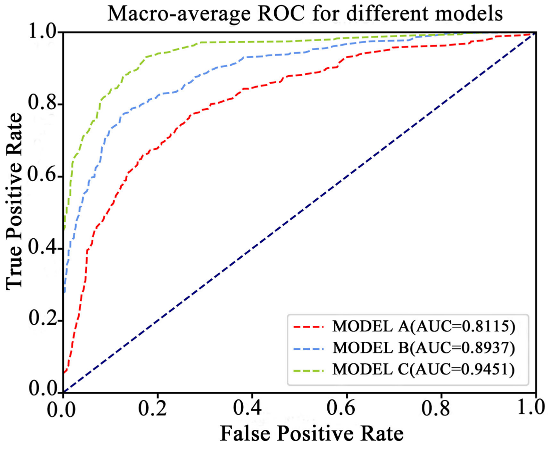

Fig. 8. ROC curves of the three models based on the test dataset after SMOTE.

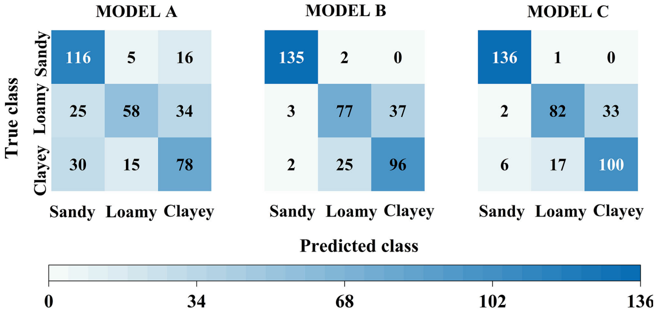

Fig. 9. Confusion matrices of the three models based on the test dataset after SMOTE.

There were discrepancies in the identification of a single soil texture class for different models. MODEL B and MODEL C with Sentinel-2 data had more advantages in recognizing clayey and loamy soils. This indicates the potential of Sentinel-2 data for classifying soil texture classes. The contributions of Sentinel-2 data to explain soil texture class variability were about 17%, 41%, and 28% for sandy, loamy, and clayey textural soils, respectively.

B. Predictor Importance

The best model (MODEL C) was then employed to explore the importance of individual variables for identifying soil texture classes with SHAP technique. Fig. 10 shows the summary map of the variable importance based on the average absolute value of SHAP. Figs. 11–13 present SHAP summary plots for sandy, loamy, and clayey textural classes, respectively.

Overall, elevation and stratum were the main explanatory variables for soil texture classification (see Fig. 10). Band 5 in September (B5_Sep) and band 7 in October (B7_Oct) in red-edge bands also played an essential role with the mean absolute values of SHAP of 0.140 and 0.136, respectively. Among the traditional spectral indices and bands, NDWI in December (NDWI_ Dec) and band 12 in April (B12_Apr) had higher contributions to texture classification. However, these six indicators did not contribute significantly to identify each single soil texture class. According the absolute mean values of

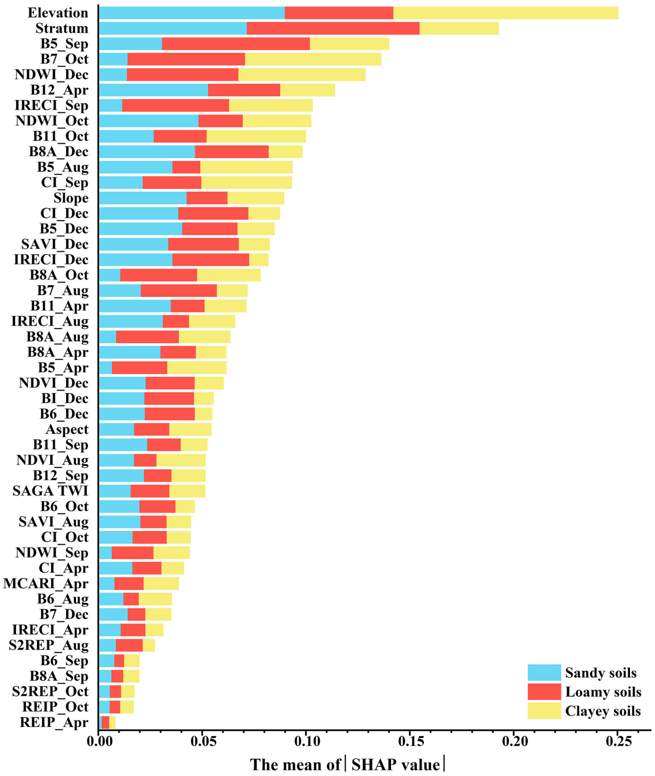

Fig. 10. Variable importance to soil texture classes (Apr., Aug., Oct., and Dec. mean April, August, October, and December, respectively).

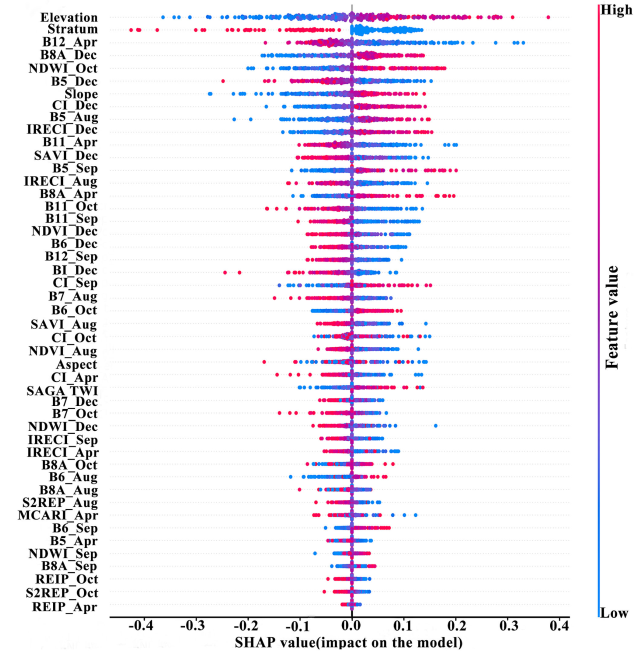

Fig. 11. SHAP summary plot of sandy soils (Each dot corresponds to a soil instance from test data. And the gradient color of dots reflects the variable's value changing from low (blue) to high (red). The input variables are placed along y-axis based on their importance, and the most influential variables are kept at the top. The x-axis represents the SHAP value. The SHAP value denotes the probability of being predicted as target textural class. The higher the SHAP value, the higher the probability. For stratum, 0 and 1 denote the Daye Formation and the Xujiahe Formation, respectively, Figs. 12–13 are same).

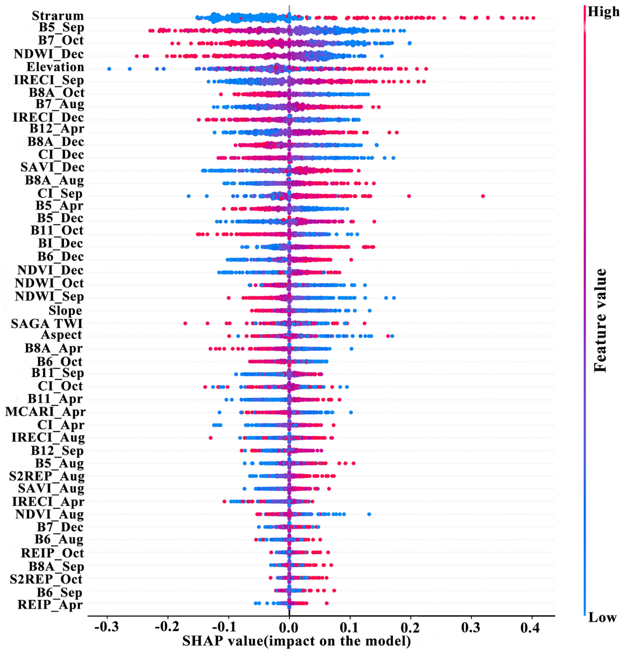

Fig. 12. SHAP summary plot of loamy soil.

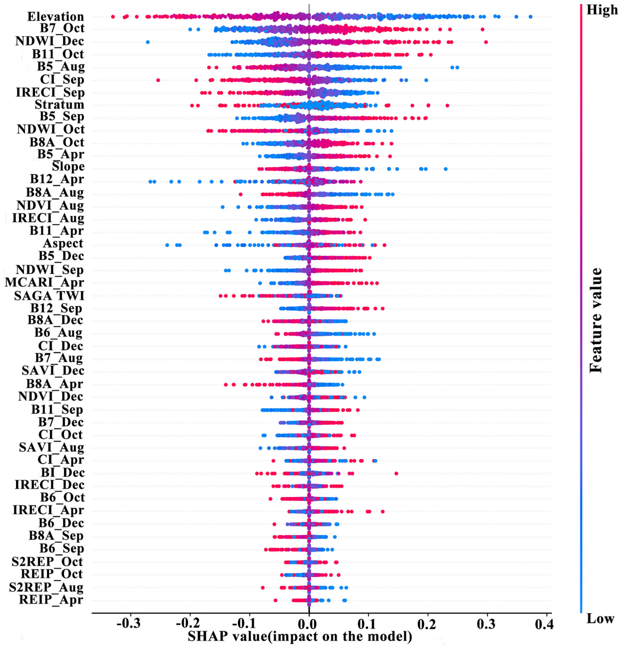

Fig. 13. SHAP summary plot of clayey soils.

SHAP, the variables with higher relative importance scores were elevation (100%) > stratum (80%) > B12_Apr (54%) for sandy soils, stratum (100%) > B5_Sep (86%) > B7_Oct (68%) > NDWI_ Dec (64%) for loamy soils, elevation (100%) > B7_Oct (60%) > NDWI_ Dec (57%) for clayey soils.

Furthermore, how the variables drive the output of the model could be revealed through summary plots generated by SHAP for understanding the decision process of the ML model [33], [68]. Figs. 11–13 show that the dots of variables with lower importance were mostly stacked vertically near the *y*-axis (SHAP value is close 0) at the bottom of the panel, while the dots of variables with higher importance tended to have wide distribution range along the x-axis. In addition, the top six variables explaining the overall variability of soil texture classes were also located in the higher parts of the panels for sandy (see Fig. 11), loamy (see Fig. 12), and clayey (see Fig. 13) textures. For example, elevation could discern sandy and clayey soils. As the elevation increased (the dot color transition from blue to red), the probability of being sandy texture increased (SHAP values change from negative to positive, see Fig. 11) and the probability of being clayey texture decreased (SHAP value changes from positive to negative, see Fig. 13). This indicated that most of the sandy soils were located at a higher elevation and clayey soils at lower areas. Stratum was useful for distinguishing loamy and sandy soils. The Xujiahe Formation (red dots) corresponded to the lower potential of sandy textural class (SHAP value < 0, see Fig. 11) and the higher tendency of loamy textural class (SHAP value > 0, see Fig. 12), showing that soils developed from the Xujiahe Formation had more loamy soils than the sandy soils. Similarly, the contributions of B7_Oct and NDWI _ Dec were mainly to separate loamy soils from clayey soils. Their effects on the output of the model could be summarized as a higher possibility of being clayey class (SHAP value > 0, see Fig. 13) and a lower possibility of being loamy class (SHAP value < 0, see Fig. 12) with the increase of their values (from blue to red). This suggested that loamy textural soils tended to had lower values of B7_Oct and clayey textural soils tended to have higher values of NDWI_Dec. B5_Sep ranked second in identifying loamy texture (see Fig. 12), and B12_Apr was third in recognizing sandy texture (see Fig. 11). Their lower values (blue dots) resulted in lower likelihood (SHAP value $<$ 0, see Figs. 11 and 12) of being loamy or sandy textual class, respectively. This pointed out that loamy soils had a lower value of B5_Sep and sandy soils had lower values of B12_Apr.

C. Soil Texture Classification Map

The soil texture class of each grid (10×10 m) for dry farmland was predicted based on the best model (MODEL C) over the entire study area (see Fig. 14). The spatial pattern of the soil texture class was in consistent with that of the stratum and elevation. Sandy soils were scattered in the northeast and southwest areas, where the dominate soils were developed from the Daye Formation at higher elevations. Loamy soils were mostly distributed in the northern part, where soils were formed from the Xujiahe Formation. Clayey soils were mainly concentrated in the west and south areas, where the elevation was lower. In addition, about 6% of the total dry farmland was sandy soils, 19% was clayey soils and 76% was loamy soils.

IV. Discussion

A. Identification of Soil Texture Classes Using Sentinel-2 Data

This study provided new insights for the potential of Sentinel-2 in DSM by using retrieved vegetation properties as proxies for soil properties. Multispectral data can not only capture directly bare soil reflectance but also obtain indirectly vegetation information to predict soil properties. For temperate agroecosystems, the high revisit rate of Sentinel-2 (providing an image every 5 days) increases the possibility of acquiring bare soils images along the crop cycle [70]. Bare soil areas could be maximized by aggregating multiple acquisition dates [70]. However, it is not applicable to our study area. The powerful capability of Sentinel-2 in the vegetation remote sensing community offers more potential usability for this study [71], [72]. The class of soil texture affects the availability of soil moisture, heat capacity, and other soil properties, which in turn leads to the response of vegetation characteristics [9]. The new spectral channels of the Sentinel-2, such as the three red-edge bands (Bands 5–7) that may be overlooked by traditional multispectral satellites, are promising in terms of detecting vegetation properties [71], [72]. Richer bands of Sentinel-2 could provide invaluable information about vegetation cover to characterize vegetation-soil response, which may be sufficient to discrete different vegetation characteristics and their associated soil attribute (soil texture class). Wang *et al.* [73] reported that variables related to vegetation were important and practical for predicting soil properties. Zhang *et al.* [25] emphasized that if there were no suitable images under bare soil conditions, more predictors should be considered to improve model performance under vegetation condition, for which Sentinel-2 may be an appropriate data source.

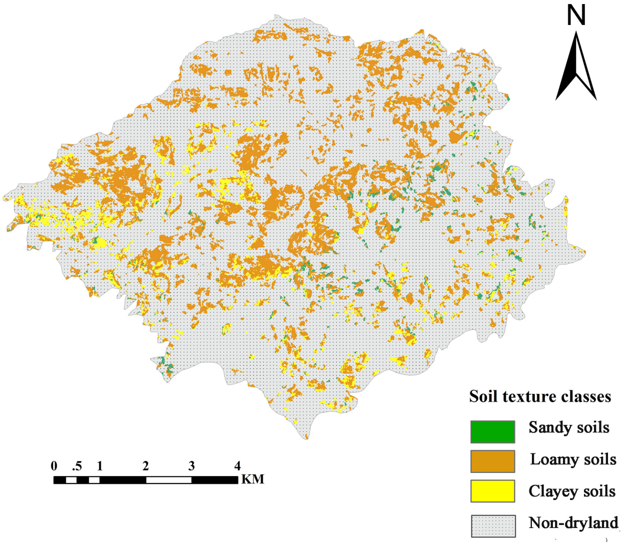

Fig. 14. Soil texture classification map.

We also found that the prediction accuracy of the model with purely Sentinel-2 data is superior to the model with solely DEM derivatives and stratum (see Fig. 7). Enough heterogeneity of vegetation features captured by Sentinle-2 might involve more information about the target soil properties, because the vegetation is usually a comprehensive expression of several factors, such as soils and terrain [9]. In addition, the combination of DEM, stratum and Sentinel-2 data produced the most accurate classification result, proving that both of them have unique explanatory power for soil properties prediction.

Compared with existing studies, the results of identifying soil texture class based on Sentinel-2 data in this study are better than those based on Landsat series data (OA < 0.67, kappa < 0.53) [11], [74]. Higher spatial resolution seems to be critical in the prediction of spatial variability of soil properties. Ceddia *et al.* [16] pointed out that coarser spectral resolution data (e.g., Landsat 8 OLI with 30 m resolution) are more likely to have mixed-pixel problems and hence less sensitivity to spatial complexity at short distances. The higher spectral resolution may capture high levels of detail in landscape scale and smaller objects (e.g., individual vegetation) whose spectral behaviors are highly variable [75]. Sentinel-2 with high spectral resolution (10 or 20 m) may better represent the spatial variability of soil texture classes in detail.

B. Critical Factors Related To Soil Texture Classes Based on the SHAP

The global importance of all variables was shown by using the SHAP method. We found that elevation and stratum were the most momentous predictors, followed by red-edge bands (B5 in September and B7 in October), NDWI in December, and B12 in April (see Fig. 10). Furthermore, the direction of the impacts and contribution degree of these six variables (relative score $> 50\%$) for different soil texture classes identification were revealed (see Figs. 11–13).

Elevation plays a key role in the development of microclimates, and in turn affects the soil processes [76]. It determines gravity and hydraulic power conditions and thus the intensity of erosion, redistribution, and sorting processes of soil particles to a large extent [2]. Wilcke *et al.* [77] noted a strong dependence of soil texture on elevation. Specifically, they pointed out a good positive correlation between elevation and sand content but a negative correlation between elevation and clay content, which is consistent with our findings (see Figs. 11 and 13). This might be attributed to the down-profile and downslope removal of finer particles [78]. In previous studies of soil properties predictions, elevation was also found to be the most effective topographic parameter [2], [3], [77]. Geological factors exert a strong control on potential pedogenesis at larger scales [79]. The classes of soil texture developed from different strata may be diverse. Soils developed from the Xujiahe Formation having a more loamy texture than sandy texture (see Figs. 11 and 12) might be resulted from the mudstone composition of this stratum in our study area. Fityus and Smith [80] also reported that mudstone and marls usually formed fine-textured soils.

Soil texture plays a key role in soil functions, including fertility and solute transport [81]. Hence, there are differences in the growth of crops for soils with distinct texture classes. In September and October, the farmlands are shaded by sweet potato (root crops) leaves. The growth state of vegetation leaves could reflect the sensitivity of root crops to soil textual classes. Bands 7 and 5, which are closely related to biophysical and chemical properties of vegetation (e.g., plant nitrogen uptake) [25], [82], may characterize the difference in crop leaves under diverse soil texture classes. In addition, the foliar spectral reflectance decreases with the increase of vegetation vitality between 500 and 740 nm (central wavelength of band 5 is 703 nm) [83]. Thus, the lower reflectance of B5 (in September) in loamy soils than sandy and clayey soils (see Figs. 11–13) reflected the fact that soils with sandy loam texture are better for tuber and root crops [84]. Same as shown in Fig. 11, Gholizadeh *et al.* [48] also found that band 7 of Sentinel-2 images provided a good positive correlation with clayey soils under bare soils. Therefore, the usefulness of the red-edge factors of Sentinel-2 data in predicting soil properties are worthy to be explored in the future.

In addition, plenty of references reported the close relationship between soil texture and soil moisture [81], [85]. The available soil water content depends greatly on the soil texture [86]. Soil moisture affects the intensity and absorption features of spectral reflectance, causing the difference in spectral reflectance under different soil texture classes. Band 12 (SWIR) and NDWI that are related to soil moisture are important variables for predicting soil texture in our study [87], [88]. Liao *et al.* [86] found a good positive correlation between the reflectance of SWIR and sand ($r = 0.57$, $p < 0.01$), which is inconsistent with ours (see Fig. 11). However, their finding is on bare soil (NDVI < 0.1). In fact, the spectral reflectance captured by the satellite-borne sensor is also affected by soil roughness, climatic conditions, and vegetation cover in addition to soil moisture [81], [89]. Our study area is planted with crops, where the spectral reflectance might be related to vegetation cover to a large extent. Crop water budget changes are sensitive to soil texture under climatic changes [90]. Therefore, the importance of band 12 and NDWI might be ascribed to the spectral response of vegetation moisture [91], [92]. Specifically, the soil with high water content and low porosity suppressed the progress of soil temperature decreasing in Winter and increasing in Spring [93]. Then, soil temperature influences biological processes, like the uptake of water and nutrients by roots [94]. Therefore, crop moisture content might be higher in clayey soils in Winter (NDWI_Dec, see Fig. 13), and higher in sandy soils in Spring (B12_Apr, see Fig. 11).

C. Deficiencies and Prospects

In fact, few studies have predicted soil texture classes with Sentinel-2, and the results of them do not seem to be ideal (OA ranging from 0.43 to 0.65, kappa ranging 0.20 from 0.46) in comparison with the current work [47], [81]. The difference in prediction accuracy might be associated with the conditions of the research area (e.g., vegetation cover, soil, and climate), the models and sampling data. The current study was conducted under one land use type, and the samples did not cover all soil texture classes (e.g., no silty soils). This limitation suggested the research might be regional and the model might be applied in other areas with similar conditions. The application potential of Sentinel-2 in DSM needs to be further proved in larger areas with more soil texture classes under different land use types.

Additionally, how to improve the accuracy of mapping soil texture deserves further exploration. DSM might benefit from the synthetic aperture radar (SAR) data, such as Sentinel-1, due to their all-weather, day, and night imaging advantages [76], [81]. They can provide information beyond the vegetative cover and the soil surface, and broaden the potential of soil properties characterization in areas where optical satellite sensors cannot observe the ground. Besides, some new strategies proposed may grant the great prospect for DSM, such as deep learning (DL) algorithms [95], [96], and ensemble learning modelings [97]. Compared with traditional ML, they have unique advantages. DL learns data itself to automatically capture high-level features, avoiding the subjectivity of handcrafted feature selection in ML [98]. Ensemble strategy can overcome the disadvantages of each individual ML to improve model performance [99]. Future research will focus on the conjunction of SAR data and multitemporal optical images with different methods to improve classification accuracy.

V. Conclusion

In this study, based on the close correlation between environment variables and soil properties, the potential of Sentinel-2 in the identification of soil texture classes was explored. The SHAP method was used to visually display the weights of different variables and their relationships with the target. The main findings are summarized as follows:

1) SVM with the combination of Sentinel-2 data, terrain, and stratum achieved the highest classification accuracy, and that with purely Sentinel-2 data was also good in soil texture classification. This confirmed the new potential of Sentinel-2 in predicting soil properties.
2) The ranking of the variable importance indicated that elevation, stratum, band 5 in September and band 7 in October, NDWI in December, and band 12 in April were the key predictors. In particular, the red-edge factors are worthy of further study in DSM.
3) SHAP method showed both global and local contributions of variables to soil texture classes. It also revealed how the changes in variable values affected the final prediction direction. This provided promising technical support for improving the interpretability of ML in DSM tasks.

Appendix

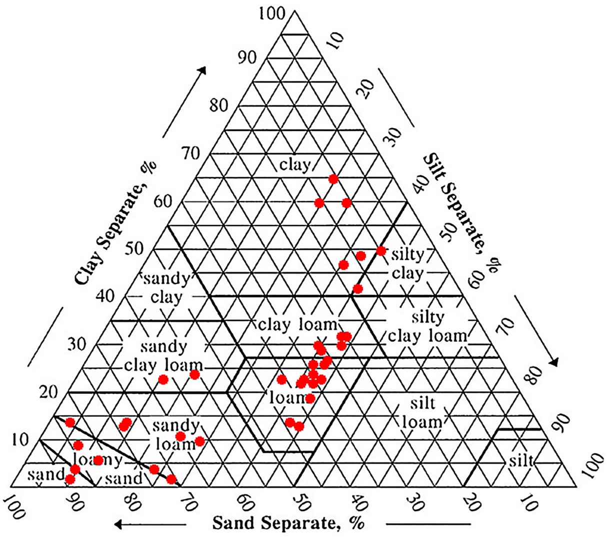

Fig. 15. Soil texture classification of 43 samples based on laboratory analysis.

TABLE IV
General Soil Texture Groups Classification Confusion Matrix and Accuracy Evaluation Indicators

	Sandy soils	Loamy soils	Clayey soils	Kappa	OA
Sandy soils	6	1	0	0.604	0.767
Loamy soils	3	20	6		
Clayey soils	0	0	7		

TABLE V
A List of Simple Count for the 42 Remote Sensing Variables

Time phase	Remote sensing variable
August 29th,2018	B5, B7, IRECI, B8A, NDVI, SAVI, B6, S2REP
September 20th,2018	B5, IRECI, CI, B11, B12, NDWI, B6, B8A
October 18th,2018	B7, NDWI, B11, B8A, B6, CI, REIP, S2REP
April 16th,2019	B12, B11, B8A, B5, CI, MCARI, IRECI, REIP
December 12th,2019	NDWI, B8A, CI, B5, SAVI, IRECI, NDVI, BI, B6, B7

References

[1] L. Poggio and A. Gimona, “3D mapping of soil texture in Scotland,” *Geoderma Reg.*, vol. 9, pp. 5–16, 2017.

[2] F. Liu *et al.*, “High-resolution and three-dimensional mapping of soil texture of China,” *Geoderma*, vol. 361, 2020, Art. no. 114061.

[3] M. Ließ, B. Glaser, and B. Huwe, “Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and random forest models,” *Geoderma*, vol. 170, pp. 70–79, 2012.

[4] V. L. Mulder, M. Lacoste, A. C. Richer-de-Forges, and D. Arrouays, “GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth,” *Sci. Total Environ.*, vol. 573, pp. 1352–1369, 2016.

[5] J. Padarian, B. Minasny, and A. B. McBratney, “Chile and the Chilean soil grid: A contribution to globalsoilmap,” *Geoderma Reg.*, vol. 9, pp. 17–28, 2017.

[6] Z. Bian *et al.*, “Applying statistical methods to map soil organic carbon of agricultural lands in northeastern coastal areas of China,” *Arch. Agron. Soil Sci.*, pp. 1–13, 2019.

[7] R. Casa, F. Castaldi, S. Pascucci, A. Palombo, and S. Pignatti, “A comparison of sensor resolution and calibration strategies for soil texture estimation from hyperspectral remote sensing,” *Geoderma*, vol. 197–198, pp. 17–26, 2013.

[8] E. Jalilvand, M. Tajrishy, S. A. Ghazi Zadeh Hashemi, and L. Brocca, “Quantification of irrigation water using remote sensing of soil moisture in a semi-arid region,” *Remote Sens. Environ.*, vol. 231, 2019, Art. no. 111226.

[9] J. J. Maynard and M. R. Levi, “Hyper-temporal remote sensing for digital soil mapping: Characterizing soil-vegetation response to climatic variability,” *Geoderma*, vol. 285, pp. 94–109, 2017.

[10] J. G. Kalambukattu, S. Kumar, and R. Arya Raj, “Digital soil mapping in a Himalayan watershed using remote sensing and terrain parameters employing artificial neural network model,” *Environ. Earth Sci.*, vol. 77, no. 5, 2018, Art. no. 203.

[11] Y. Zhai, J. A. Thomasson, J. E. Boggess, and R. Sui, “Soil texture classification with artificial neural networks operating on remote sensing data,” *Comput. Electron. Agric.*, vol. 54, no. 2, pp. 53–68, 2006.

[12] C. da Silva Chagas, W. de Carvalho Junior, S. B. Bhering, and B. Calderano Filho, “Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions,” *Catena*, vol. 139, pp. 232–240, 2016.

[13] B. C. Gallo *et al.*, “Multi-temporal satellite images on topsoil attribute quantification and the relationship with soil classes and geology,” *Remote Sens.*, vol. 10, no. 10, 2018, Art. no. 1571.

[14] A. Gasmi, C. Gomez, P. Lagacherie, H. Zouari, A. Laamrani, and A. Chehbouni, “Mean spectral reflectance from bare soil pixels along a Landsat-TM time series to increase both the prediction accuracy of soil clay content and mapping coverage,” *Geoderma*, vol. 388, 2021, Art. no. 114864.

[15] T. Loiseau *et al.*, "Satellite data integration for soil clay content modelling at a national scale," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 82, 2019, Art. no. 101905.

[16] M. Ceddia, A. Gomes, G. Vasques, and É. Pinheiro, "Soil carbon stock and particle size fractions in the central amazon predicted from remotely sensed relief, multispectral and radar data," *Remote Sens.*, vol. 9, no. 2, 2017, Art. no. 124.

[17] J. A. M. Demattê, V. M. Sayão, R. Rizzo, and C. T. Fongaro, "Soil class and attribute dynamics and their relationship with natural vegetation based on satellite remote sensing," *Geoderma*, vol. 302, pp. 39–51, 2017.

[18] W. de Carvalho Junior, P. Lagacherie, C. da Silva Chagas, B. Calderano Filho, and S. B. Bhering, "A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment," *Geoderma*, vol. 232, pp. 479–486, 2014.

[19] E. Grabska, D. Frantz, and K. Ostapowicz, "Evaluation of machine learning algorithms for forest stand species mapping using sentinel-2 imagery and environmental data in the Polish carpathians," *Remote Sens. Environ.*, vol. 251, 2020, Art. no. 112103.

[20] M. Persson, E. Lindberg, and H. Reese, "Tree species classification with multi-temporal sentinel-2 data," *Remote Sens.*, vol. 10, no. 11, 2018, Art. no. 1794.

[21] H. Zhang, J. Kang, X. Xu, and L. Zhang, "Accessing the temporal and spectral features in crop type mapping using multi-temporal sentinel-2 imagery: A case study of Yi'an county, Heilongjiang province, China," *Comput. Electron. Agric.*, vol. 176, 2020, Art. no. 105618.

[22] A. B. McBratney, M. L. Mendonça Santos, and B. Minasny, "On digital soil mapping," *Geoderma*, vol. 117, no. 1/2, pp. 3–52, 2003.

[23] C. Wang, S. Wang, B. Fu, Z. Li, X. Wu, and Q. Tang, "Precipitation gradient determines the tradeoff between soil moisture and soil organic carbon, total nitrogen, and species richness in the loess plateau, China," *Sci. Total Environ.*, vol. 575, pp. 1538–1545, 2017.

[24] Y. Xu, S. E. Smith, S. Grunwald, A. Abd-Elrahman, S. P. Wani, and V. D. Nair, "Estimating soil total nitrogen in smallholder farm settings using remote sensing spectral indices and regression kriging," *Catena*, vol. 163, pp. 111–122, 2018.

[25] Y. Zhang, B. Sui, H. Shen, and L. Ouyang, "Mapping stocks of soil total nitrogen using remote sensing data: A comparison of random forest models with different predictors," *Comput. Electron. Agric.*, vol. 160, pp. 23–30, 2019.

[26] W. Wu, Q. Yang, J. Lv, A. Li, and H. Liu, "Investigation of remote sensing imageries for identifying soil texture classes using classification methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1653–1663, Mar. 2019.

[27] M. A. Ghorbani, S. Shamshirband, D. Zare Haghi, A. Azani, H. Bonakdari, and I. Ebtehaj, "Application of firefly algorithm-based support vector machines for prediction of field capacity and permanent wilting point," *Soil Tillage Res.*, vol. 172, pp. 32–38, 2017.

[28] L. Deiss, A. J. Margenot, S. W. Culman, and M. S. Demyan, "Tuning support vector machines regression models improves prediction accuracy of soil properties in MIR spectroscopy," *Geoderma*, vol. 365, 2020, Art. no. 114227.

[29] L. Zhang, Y. Wang, M. Niu, C. Wang, and Z. Wang, "Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan rural cohort study," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020.

[30] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018.

[31] A. Stojić *et al.*, "Explainable extreme gradient boosting tree-based prediction of toluene, ethylbenzene and xylene wet deposition," *Sci. Total Environ.*, vol. 653, pp. 140–147, 2019.

[32] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, pp. 4765–4774, 2017.

[33] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," 2018, *arXiv1802.03888*.

[34] Y. Guo *et al.*, "Effects of microplastics on growth, phenanthrene stress, and lipid accumulation in a diatom, phaeodactylum tricornutum," *Environ. Pollut.*, vol. 257, 2020, Art. no. 113628.

[35] R. Rodríguez-Pérez and J. Bajorath, "Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values," *J. Med. Chem.*, vol. 63, pp. 8761–8777, 2019.

[36] J. Xu, M. Saleh, and M. Hatzopoulou, "A machine learning approach capturing the effects of driving behaviour and driver characteristics on trip-level emissions," *Atmos. Environ.*, vol. 224, 2020, Art. no. 117311.

[37] F. Akhtar *et al.*, "Diagnosis and prediction of large-for-gestational-age fetus using the stacked generalization method," *Appl. Sci.*, vol. 9, no. 20, 2019, Art. no. 4317.

[38] Z. T. Gong, *Chinese Soil Taxonomy (in Chinese)*. Beijing, China: Sci. Press, 1999.

[39] F. A. O. FAO, "Unesco soil map of the world, revised legend, with corrections and updates," *World Soil Resour. Rep.*, vol. 60, 1988, Art. no. 140.

[40] D. F. Post, A. R. Huete, and D. S. Pease, "A comparison of soil scientist estimations and laboratory determinations of some Arizona soil properties," *J. Soil Water Conserv.*, vol. 41, no. 6, pp. 421–424, 1986.

[41] C. Vos, A. Don, R. Prietz, A. Heidkamp, and A. Freibauer, "Field-based soil-texture estimates could replace laboratory analysis," *Geoderma*, vol. 267, pp. 215–219, 2016.

[42] Y. A. Pachepsky, W. J. Rawls, and H. S. Lin, "Hydropedology and pedotransfer functions," *Geoderma*, vol. 131, no. 3/4, pp. 308–316, 2006.

[43] W. Wu, A.-D. Li, X.-H. He, R. Ma, H.-B. Liu, and J.-K. Lv, "A comparison of support vector machines, artificial neural network and classification tree for identifying soil texture classes in southwest China," *Comput. Electron. Agric.*, vol. 144, pp. 86–93, 2018.

[44] R. Jahn, H. P. Blume, V. B. Asio, O. Spaargaren, and P. Schad, "Guidelines for soil description," FAO, Rome, Italy, 2006.

[45] S. S. D. Staff, "Soil survey manual," in *USDA Handbook 18*. Govern. Printing Office, WA, DC, USA, 2017, Art. no. 639.

[46] Y. Liu, W. Gong, Y. Xing, X. Hu, and J. Gong, "Estimation of the forest stand mean height and aboveground biomass in northeast China using SAR Sentinel-1B, multispectral Sentinel-2A, and DEM imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 151, pp. 277–289, 2019.

[47] C. Gomez, S. Dharumarajan, J.-B. Féret, P. Lagacherie, L. Ruiz, and M. Sekhar, "Use of sentinel-2 time-series images for classification and uncertainty analysis of inherent biophysical property: Case of soil texture mapping," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 565.

[48] A. Gholizadeh, D. Žižala, M. Saberioon, and L. Borůvka, "Soil organic carbon and texture retrieving and mapping using proximal, airborne and sentinel-2 spectral imaging," *Remote Sens. Environ.*, vol. 218, pp. 89–103, 2018.

[49] W. Wu, Q. Yang, J. Lv, A. Li, and H. Liu, "Investigation of remote sensing imageries for identifying soil texture classes using classification methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1653–1663, Mar. 2019.

[50] J. U. H. Eitel *et al.*, "Broadband, red-edge information from satellites improves early stress detection in a New Mexico conifer woodland," *Remote Sens. Environ.*, vol. 115, no. 12, pp. 3640–3646, 2011.

[51] J. G. P. W. Clevers and A. A. Gitelson, "Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on sentinel-2 and -3," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 23, pp. 344–351, 2013.

[52] G. Guyot, F. Baret, and D. J. Major, "High spectral resolution: Determination of spectral shifts between the red and near infrared BT - ISPRS Congress," *Int. Arch. Photogramm. Remote Sens.*, vol. 11, pp. 750–760, 1988.

[53] C. Wu, Z. Niu, Q. Tang, and W. Huang, "Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation," *Agric. For. Meteorol.*, vol. 148, no. 8–9, pp. 1230–1241, 2008.

[54] A. Gitelson and M. N. Merzlyak, "Spectral reflectance changes associated with autumn senescence of aesculus hippocastanum l and acer platanoides l leaves. Spectral features and relation to chlorophyll estimation," *J. Plant Physiol.*, vol. 143, no. 3, pp. 286–292, 1994.

[55] N. L. Tsakiridis, J. B. Theocharis, E. Ben-Dor, and G. C. Zalidis, "Using interpretable fuzzy rule-based models for the estimation of soil organic carbon from VNIR/SWIR spectra and soil texture," *Chemom. Intell. Lab. Syst.*, vol. 189, pp. 39–55, 2019.

[56] P. Ceccato, S. Flasse, S. Tarantola, S. Jacquemoud, and J.-M. Grégoire, "Detecting vegetation leaf water content using reflectance in the optical domain," *Remote Sens. Environ.*, vol. 77, no. 1, pp. 22–33, 2001.

[57] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance," *Inf. Sci. (Ny).*, vol. 505, pp. 32–64, 2019.

[58] M. Ijaz, G. Alfian, M. Syafrudin, and J. Rhee, "Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest," *Appl. Sci.*, vol. 8, no. 8, 2018, Art. no. 1325.

[59] A. Niu, B. Cai, and S. Cai, "Big data analytics for complex credit risk assessment of network lending based on SMOTE algorithm," *Complexity*, vol. 2020, pp. 1–9, 2020.

[60] R. Taghizadeh-Mehrjardi *et al.*, "Synthetic resampling strategies and machine learning for digital soil mapping in Iran," *Eur. J. Soil Sci.*, vol. 71, no. 3, pp. 352–368, 2020.

[61] G. Kovács, "Smote-variants: A python implementation of 85 minority oversampling techniques," *Neurocomputing*, vol. 366, pp. 352–354, 2019.

[62] G. Louppe, "Understanding random forests: From theory to practice," 2014, *arXiv:1407.7502*.

[63] M. Kuhn and K. Johnson, "Resampling techniques," in *Applied Predictive Modeling*, New York, NY, USA: Springer, 2013, vol. 26, pp. 69–73.

[64] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

[65] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*. Hoboken, NJ, USA: Wiley, 2011.

[66] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, 2014.

[67] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," 2016, *arXiv:1606.05386*.

[68] P. Arjunan, K. Poolla, and C. Miller, "EnergyStar++: Towards more accurate and explanatory building energy benchmarking," *Appl. Energy*, vol. 276, 2020, Art. no. 115413.

[69] F. Yan, K. Song, Y. Liu, S. Chen, and J. Chen, "Predictions and mechanism analyses of the fatigue strength of steel based on machine learning," *J. Mater. Sci.*, vol. 55, no. 31, pp. 15334–15349, 2020.

[70] E. Vaudour, C. Gomez, Y. Fouad, and P. Lagacherie, "Sentinel-2 image capacities to predict common topsoil properties of temperate and mediterranean agroecosystems," *Remote Sens. Environ.*, vol. 223, pp. 21–33, 2019.

[71] I. Herrmann, A. Pimstein, A. Karnieli, Y. Cohen, V. Alchanatis, and D. J. Bonfil, "LAI assessment of wheat and potato crops by VENμS and sentinel-2 bands," *Remote Sens. Environ.*, vol. 115, no. 8, pp. 2141–2151, 2011.

[72] J. Delegido, J. Verrelst, L. Alonso, and J. Moreno, "Evaluation of sentinel-2 red-edge bands for empirical estimation of green LAI and chlorophyll content," *Sensors*, vol. 11, no. 7, pp. 7063–7081, 2011.

[73] S. Wang *et al.*, "Mapping total soil nitrogen from a site in northeastern China," *CATENA*, vol. 166, pp. 134–146, 2018.

[74] J. A. M. Demattê, M. R. Alves, F. da S. Terra, R. W. D. Bosquilia, C. T. Fongaro, and P. P. da S. Barros, "Is it possible to classify topsoil texture using a sensor located 800 km away from the surface?," *Rev. Bras. Ciência do Solo*, vol. 40, 2016, Art. no. e0150335.

[75] A. Samuel-Rosa, G. B. M. Heuvelink, G. M. Vasques, and L. H. C. Anjos, "Do more detailed environmental covariates deliver more accurate soil maps?," *Geoderma*, vol. 243, pp. 214–227, 2015.

[76] T. Zhou, Y. Geng, J. Chen, J. Pan, D. Haase, and A. Lausch, "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.*, vol. 729, 2020, Art. no. 138244.

[77] W. Wilcke, S. Yasin, A. Schmitt, C. Valarezo, and W. Zech, "Soils along the altitudinal transect and in catchments," in *Gradients in a Tropical Mountain Ecosystem of Ecuador*. Berlin, Germany: Springer, 2008, pp. 75–85.

[78] P. E. Gessler, O. A. Chadwick, F. Chamran, L. Althouse, and K. Holmes, "Modeling soil–landscape and ecosystem properties using terrain attributes," *Soil Sci. Soc. Amer. J.*, vol. 64, no. 6, pp. 2046–2056, 2000.

[79] C. W. Brungard, J. L. Boettinger, M. C. Duniway, S. A. Wills, and T. C. Edwards, "Machine learning for predicting soil classes in three semi-arid landscapes," *Geoderma*, vol. 239–240, pp. 68–83, 2015.

[80] S. G. Fityus and D. W. Smith, "The development of a residual soil profile from a mudstone in a temperate climate," *Eng. Geol.*, vol. 74, no. 1–2, pp. 39–56, 2004.

[81] S. Bousbih *et al.*, "Soil texture estimation using radar and optical data from sentinel-1 and Sentinel-2," *Remote Sens. (Basel, Switzerland)*, vol. 11, no. 13, 2019, Art. no. 1520.

[82] L. Korhonen, P. Packalen, H. Hadi, and M. Rautiainen, "Comparison of sentinel-2 and landsat 8 in the estimation of boreal forest canopy cover and leaf area index," *Remote Sens. Environ.*, vol. 195, pp. 259–274, 2017.

[83] J. E. Luther and A. L. Carroll, "Development of an index of balsam fir vigor by foliar spectral reflectance," *Remote Sens. Environ.*, vol. 69, no. 3, pp. 241–252, 1999.

[84] S. H. Ahmadi *et al.*, "Interaction of different irrigation strategies and soil textures on the nitrogen uptake of field grown potatoes," *Int. J. Plant Prod.*, vol. 5, pp. 263–274, 2011.

[85] F. Castaldi, A. Palombo, S. Pascucci, S. Pignatti, F. Santini, and R. Casa, "Reducing the influence of soil moisture on the estimation of clay from hyperspectral data: A case study using simulated PRISMA data," *Remote Sens.*, vol. 7, no. 11, pp. 15561–15582, 2015.

[86] K. Liao, S. Xu, J. Wu, and Q. Zhu, "Spatial estimation of surface soil texture using remote sensing data," *Soil Sci. Plant Nutr.*, vol. 59, no. 4, pp. 488–500, 2013.

[87] Y. Zhang, K. Tan, X. Wang, and Y. Chen, "Retrieval of soil moisture content based on a modified hapke photometric model: A novel method applied to laboratory hyperspectral and sentinel-2 MSI data," *Remote Sens.*, vol. 12, no. 14, 2020, Art. no. 2239.

[88] M. Hosseini and M. R. Saradjian, "Multi-index-based soil moisture estimation using MODIS images," *Int. J. Remote Sens.*, vol. 32, no. 21, pp. 6799–6809, 2011.

[89] F. Castaldi, A. Palombo, F. Santini, S. Pascucci, S. Pignatti, and R. Casa, "Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon," *Remote Sens. Environ.*, vol. 179, pp. 54–65, 2016.

[90] M. Nouri, M. Homaee, M. Bannayan, and G. Hoogenboom, "Towards modeling soil texture-specific sensitivity of wheat yield and water balance to climatic changes," *Agric. Water Manag.*, vol. 177, pp. 248–263, 2016.

[91] L. Wang, J. J. Qu, X. Hao, and Q. Zhu, "Sensitivity studies of the moisture effects on MODIS SWIR reflectance and vegetation water indices," *Int. J. Remote Sens.*, vol. 29, no. 24, pp. 7065–7075, 2008.

[92] Y. Gu, E. Hunt, B. Wardlow, J. B. Basara, J. F. Brown, and J. P. Verdin, "Evaluation of MODIS NDVI and NDWI for vegetation drought monitoring using oklahoma mesonet soil moisture data," *Geophys. Res. Lett.,* vol. 35, no. 22, 2008, Art. no. L22401.

[93] T. Arkhangelskaya and K. Lukyashchenko, "Estimating soil thermal diffusivity at different water contents from easily available data on soil texture, bulk density, and organic carbon content," *Biosyst. Eng.*, vol. 168, pp. 83–95, 2018.

[94] D. Dec, J. Dörner, and R. Horn, "Effect of soil management on their thermal properties," *J. Soil Sci. Plant Nutr.*, vol. 9, no. 1, pp. 26–39, 2009.

[95] N. Wambugu *et al.*, "Hyperspectral image classification on insufficient-sample and feature learning using deep neural networks: A review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 105, 2021, Art. no. 102603.

[96] N. Tziolas, N. Tsakiridis, E. Ben-Dor, J. Theocharis, and G. Zalidis, "Employing a multi-input deep convolutional neural network to derive soil clay content from a synergy of multi-temporal optical and radar imagery data," *Remote Sens.*, vol. 12, no. 9, 2020, Art. no. 1389.

[97] K. Tan *et al.*, "Estimating the distribution trend of soil heavy metals in mining area from hymap airborne hyperspectral imagery based on ensemble learning," *J. Hazard. Mater.*, vol. 401, 2021, Art. no. 123288.

[98] S. Liu, Q. Shi, and L. Zhang, "Few-shot hyperspectral image classification with unknown classes using multitask deep learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5085–5102, Jun. 2021.

[99] R. Taghizadeh-Mehrjardi *et al.*, "Enhancing the accuracy of machine learning models using the super learner technique in digital soil mapping," *Geoderma*, vol. 399, 2021, Art. no. 115108.

Yanan Zhou received the bachelor's degree in land resource management from Southwest University, Chongqing, China, in 2020. She is currently working toward the master's degree with the Chongqing Key Laboratory of Digital Agriculture, Chongqing, China.

Her research interests include remote sensing application and geographic information system.

Wei Wu received the Ph.D. degree in agriculture from Southwest University, Chongqing, China, in 2007.

She is currently a Professor with the College of Computer and Information Science, Southwest University. Her research interests include computational intelligence with applications to decision support, data mining, and image understanding.

Huan Wang, photograph and biography not available at the time of publication.

Xin Zhang, photograph and biography not available at the time of publication.

Chao Yang, photograph and biography not available at the time of publication.

Hongbin Liu received the Ph.D. degree in utilization science of agriculture resources from Southwest University, Chongqing, China, in 2002.

He is currently a Professor with College of Resources and Environment, Southwest University. His research interests include the impacts of environment factors, such as soil, climate, on crop growth and topography, and the development of predictive models for improving the understanding of the relationships between topography and soil properties based on the technologies of geographical information system and remote sensing.