remote sensing

MDPI

Article

High-Accuracy Mapping of Soil Parent Material Types in Hilly Areas at the County Scale Using Machine Learning Algorithms

Xueliang Zeng [1,2], Xi Guo [1,2,*], Yefeng Jiang [1,2], Weifeng Li [1,2], Jiaxin Guo [1,2], Qiqing Zhou [1,2] and Hengyu Zou [1,2]

[1] College of Land Resources and Environment, Jiangxi Agricultural University, Nanchang 330045, China; zengxl307@163.com
[2] Key Laboratory of Poyang Lake Watershed Agricultural Resources and Ecology of Jiangxi Province, Nanchang 330045, China
* Correspondence: guoxi@jxau.edu.cn

Abstract: Conventional maps of soil parent material (SPM) types obtained by field survey and manual mapping or predictions from other map data have limited accuracy. Digital soil mapping of SPM types necessitates accurate acquisition of SPM distribution information, which is still a challenge in hilly areas. This study developed a high-accuracy method for SPM identification in hilly areas at the county scale. Based on geographic information system technology, seven feature variables were extracted from the geological map, geomorphic map, digital elevation model, and remote sensing image data of Shanggao County, Jiangxi Province, China. Different feature combination schemes were designed to develop SPM identification models based on random forest (RF), support vector machine (SVM), and maximum likelihood classification (MLC) algorithms. The best SPM identification results were obtained from the RF algorithm using the combination of geological type, geomorphic type, elevation, and slope. Confusion matrices were constructed based on a field survey of 586 validation samples, and the results were evaluated in terms of overall accuracy, precision, recall, F1 score, and Kappa coefficient. The overall accuracy and Kappa coefficient of the results from the optimal RF model were 83.11% and 0.79, respectively, which were 26.11% and 0.31 higher than those of the conventional map, respectively. Its precision and recall for various SPM types were greater than 75%. A comprehensive comparison of the accuracy, uncertainty, and plotting performance of the SPM recognition results reveals that the RF algorithm outperforms the SVM algorithm and the MLC algorithm. Geological type was the largest contributor to SPM identification, followed by geomorphic type, elevation, and slope. The importance of different feature variables varied for distinct SPM types. The accuracy of SPM identification was not improved by selecting more feature variables, such as land use type, normalised difference vegetation index, and topographic wetness index. This study demonstrates the feasibility of high-accuracy county-level SPM mapping in hilly areas based on the RF algorithm using geological type, geomorphic type, elevation, and slope as feature variables. As hilly areas have typical topographic features and SPM types, the proposed method of SPM mapping can be useful for application in other similar areas. There are a few limitations in this study with regard to data quality and resolution, feature variable selection, classification algorithm generalisation, and study area representativeness, which may affect the outcomes and need to be solved.

Keywords: soil parent material; machine learning; random forest; support vector machine; maximum likelihood classification; digital soil mapping

Citation: Zeng, X.; Guo, X.; Jiang, Y.; Li, W.; Guo, J.; Zhou, Q.; Zou, H. High-Accuracy Mapping of Soil Parent Material Types in Hilly Areas at the County Scale Using Machine Learning Algorithms. *Remote Sens.* **2024**, *16*, 91. https://doi.org/10.3390/rs16010091

Academic Editor: Thomas Alexandridis

Received: 6 November 2023
Revised: 13 December 2023
Accepted: 18 December 2023
Published: 25 December 2023

1. Introduction

As a fundamental natural resource on the Earth's surface, soil plays a crucial role in human production and living activities. It is accepted that soil results from the combined action of five factors—parent material, climate, biology, topography, and time [1]. Soil parent material (SPM), which provides the material basis for soil formation, often relates to

soil type and physicochemical properties. Thus, soil properties, such as texture, thickness, and rock content, can be characterised using SPM types [2]. Creating SPM-type maps with higher accuracy is a crucial step in digital soil mapping [3].

Conventional SPM-type maps are usually coarse in scale, and high-resolution digital maps of SPM types are rare [2,4]. For example, China only has a 1:14,000,000 SPM-type map and a 1:6,000,000 soil parent rock/material map at a nationwide scale. Heung et al. [5] obtained a 100 m resolution SPM-type map for British Columbia based on multiple environmental covariates. Mello et al. [6] obtained a SPM-type map for the Sipaulista Plateau in Brazil by multi-spectral analyses of satellite images. These digital maps can meet the application requirements of provincial and higher levels, but their accuracy is still insufficient at the county and lower levels. Therefore, it is important to decipher how to create high-accuracy county-level SPM-type maps by digital mapping techniques [7], especially in hilly areas with complicated distribution of SPM types.

SPM information is usually obtained by field surveys and manual mapping of SPM types, prediction from soil type maps, or replacing SPM distribution maps with geological maps [5–10]. Based on the field survey, soil profiles are obtained to determine SPM types; then, the boundaries of SPM types are delineated through a combination of topography, geology, and remote sensing images. While requiring high costs, this conventional method cannot guarantee map accuracy. Predicting SPM types from soil type maps is currently the most popular method in soil research. However, the resulting maps usually contain many multi-component units and lack spatially explicit representations of soil classes and their properties [9,11]. As far as China is concerned, the soil type maps in use are mainly based on the Second National Soil Survey conducted during the 1980s, and land use change would lead to variation in soil types in some areas [12]. Additionally, most of the soil-type maps were delineated manually with insufficient accuracy [13]. The use of geological maps instead of SPM maps is relatively convenient [14], but the relationship between geological types and SPM types is not a one-to-one correspondence [15,16]. There are cases where the same geological type corresponds to multiple SPM types in different topographic environments. Therefore, none of these methods can achieve comprehensive and accurate acquisition of SPM information, and the SPM information obtained by the different methods is partially inconsistent. In addition, the hilly areas have greater changes in topographic relief and complex environments, with greater differences in the spatial distribution of SPM types, as well as a greater variety of SPM causes, including residual sediment parent material, slope sediment parent material, and river transported material parent material, etc. [17,18], which makes the method of soil parent material type mapping more difficult than that of other areas, so it is necessary to explore a method applicable to the mapping of SPM types in the hilly areas.

Machine learning is the process by which computers automatically learn patterns from data and make predictions and decisions based on those patterns, which is an algorithmic method commonly used in soil classification and prediction research [19,20]. In contrast with conventional analytical techniques, machine learning is able to capture higher-order nonlinear relationships between variables from data iterations and avoid the omission of latent information [21,22]. This method has been widely used in high-accuracy classification and mapping, such as land cover classification [23], remote sensing image classification [24], and soil property mapping [25]. However, machine learning algorithms exemplified by random forest (RF) are rarely used in county-level identification of SPM types in hilly areas [5,26]. SPM formation is closely associated with topography, geomorphology, and geological type [11]. Topography has a prominent influence on ecological factors, including vegetation, which in turn reflects SPM characteristics. Although conventional soil type maps and topographic data have been used to extract SPM information and then identify SPM types, there are still limited studies that extract SPM information from a combination of other data sources [27–29], and it remains unclear which environmental feature variables are useful for SPM identification (Identification of different types of soil parent material).

The RF algorithm consists of multiple decision trees, and its training speed is faster compared with other machine learning algorithms, such as support vector machine (SVM) and maximum likelihood classification (MLC) [30]. It can work for missing values and deal with high-dimensional data [31]. SVM is a simple algorithm that can automatically find those support vectors with a high discriminatory ability for classification and, as such, maximise the class-to-class intervals. This algorithm shows excellent performance with few sample data [32]. The MLC algorithm determines the classification function by the mean and variance of each class, which allows the class of each object to be classified. It is particularly advantageous in terms of operation speed [33]. Despite their advantages, these machine learning algorithms require a certain number of training samples to proceed. Given the difficulty in the acquisition of SPM-type information [31], it is still challenging to use machine learning algorithms in SPM identification.

The aim of this study was to select the optimal feature combination scheme and classification algorithm for high-accuracy SPM identification in hilly areas at the county scale. The study area is in a typical hilly area in Jiangxi Province, southern China, with diverse soil types with complex SPM distribution. Therefore, the mapping results of this hilly area could be more representative than that of plain areas. A total of seven feature variables were extracted from four data sources: geological type map (1: 50,000), geomorphic type map (30 m), digital elevation model (DEM, 30 m), and remote sensing image (30 m). The feature variables were sequentially selected to obtain different combination schemes, which were used for SPM classification based on RF, SVM, and MLC algorithms. The feature combination scheme with the highest precision and the model with the best validation accuracy and mapping results were selected. This study provides guidance for county-level mapping of SPM types in hilly areas and contributes to the development of high-accuracy digital soil mapping.

2. Materials and Methods

2.1. Study Area

The study was carried out in Shanggao County, a typical hilly area of southern China (Figure 1A). Its geographic coordinates are 114°28′–115°10′E and 28°02′–28°25′N, with a total area of 1341 km^2. This county is part of the hilly and mountainous agroforestry zone located in the lower and middle reaches of the Yangtze River Basin. The major landforms are hills and low mountains with large undulations and no distinct ranges. The study area has a mid-subtropical monsoon climate, which climate is humid and warm, with abundant rainfall, and the average annual precipitation is about 1800 mm with plenty of sunshine, a long frost-free period, and four distinct seasons. The major soil types here are red soil, limestone soil, paddy soil, yellow soil, and fluvo-aquic soil, according to the Genetic Soil Classification of China.

There are no SPM-type maps of Shanggao County that had been created previously. In this study, a conventional SPM-type map of Shanggao County was derived from a soil-type map (1:50,000) based on the Second National Soil Survey and the Soil Records of Shanggao County (Figure 1B). The major SPM types in the study area were identified as river alluvium (RA), quaternary red clay (QRC), limestone weathered material (LWM), red sandstone weathered material (RSWM), granite weathered material (GWM), and mudstone weathered material (MWM). The conventional SPM-type map was coarse, and it only showed the approximate distribution of SPM types with no further details.

2.2. Research Framework

To obtain the optimal model and feature variable combination for county-level SPM identification in hilly areas, three machine learning algorithms (i.e., RF, SVM, MLC) are applied with different feature variable combination schemes. Then, field validation samples are used to assess the accuracy of SPM identification results. Considering the uncertainties in data sources, assumptions in feature extraction, and biases in the training or validation samples, the classification results may not fully reflect the actual distribution of SPM types.

Therefore, a comparative and selective approach is adopted to choose the model with the optimal accuracy and mapping performance. Additionally, Shapley additive explanations (SHAP) are used to analyse feature importance (Figure 2).

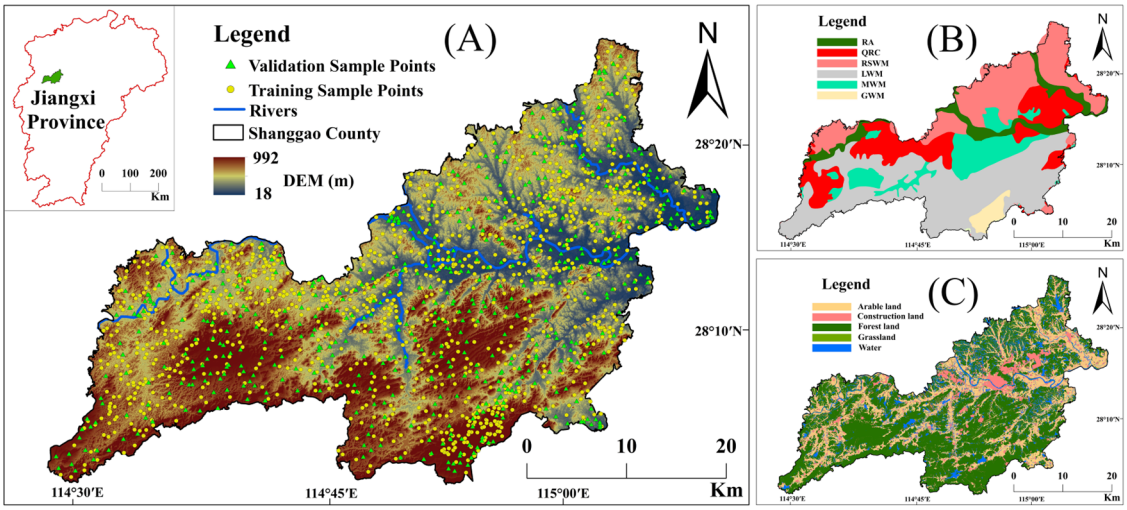

Figure 1. (**A**) Location of the sampling points in the study area (Shanggao County, Jiangxi Province, China), (**B**) conventional map of soil parent material types and (**C**) map of land use types. Digital elevation model (DEM) data were obtained in 2019, and soil parent material data were obtained in the 1980s. RA—river alluvium; QRC—Quaternary red clay; LWM—limestone weathered material; RSWM—red sandstone weathered material; GWM—granite weathered material; and MWM—mudstone weathered material. The map of land use types was identified from Landsat-8 remote sensing imagery in 2021.

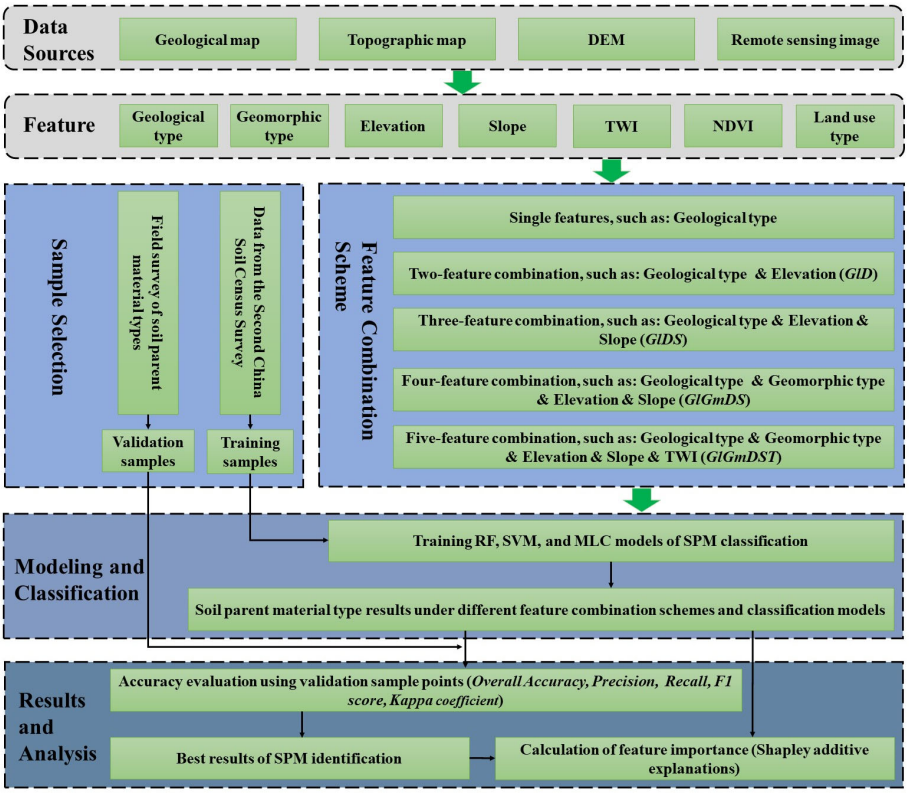

Figure 2. Research framework for machine learning-based identification of soil parent material (SPM) types in hilly areas at the county scale.

The feature combination schemes include single feature variables (e.g., geological type, geomorphic type) and combinations of two feature variables (e.g., geological type + geomorphic type, geology type + elevation), three feature variables (e.g., geological type + geomorphic type + elevation, geological type + geomorphic type + slope, geomorphic type + slope + land use type), and more feature variables. The schemes are first obtained as exhaustive combinations of two feature variables and used for modelling and SPM identification; those combinations with lower accuracy are eliminated. Then, the combinations of three or more feature variables are filtered following the same procedure.

2.3. Data Sources and Feature Extraction

As SPM results from in situ deposition or transport and accumulation of residuals after rock weathering, SPM formation is influenced by geological lithology, topography, and geomorphology [13]. In this study, geological type was extracted from the geological-type map of Shanggao County (provided by the Jiangxi Provincial Bureau of Geology, 1:50,000) and used as a geological variable; the main extraction method is classifying maps of the same lithology into one group, such as serpentine green mud kilomagnetite, metamorphic siltstone, and kilomagnetite-intercalated metamorphic sandstone into the metamorphic rock category. Elevation, slope, and topographic wetness index (TWI) [34] were extracted from DEM data (30 m) [35] as topographic variables, elevation and slope are both calculated by the spatial analysis tool in ArcGIS v10.2 software, while TWI is calculated from the unit contour length catchment area and slope, which is given by equation 1. The geomorphic type was obtained from the global basic geomorphic-type unit dataset (30 m) [36] as a geomorphic variable. The geomorphic types of Shanggao County were finally obtained as large undulating mountains, medium undulating mountains, medium undulating low mountains, low undulating low mountains, small undulating low mountains, low elevation hills, and low elevation plains.

$$TWI = \ln\left(\frac{a}{\tan\beta}\right) \quad (1)$$

where TWI is the topographic wetness index, a is the unit contour length catchment area, and β is the slope.

The presence of various SPM types leads to the development of diverse soil types, which relate to different land use types and vegetation coverage [37]. For example, SPMs of RA and sedimentary types are mainly distributed on the banks of rivers and lakes, which are suitable for agricultural cultivation and generally used as cropland. SPMs of slope and residual types are generally found in mountainous areas, dominated by forest land with high vegetation coverage. Therefore, normalised difference vegetation index (NDVI) [38] was extracted from single-phase Landsat-8 remote sensing image (30 m) [39] of June 2021 (LC08_L2SP_122040_20210119_20210307_02_T1 & LC08_L2SP_122041_20210119_20210307_02_T1) with luxuriant vegetation (Equation (2)). The RF classification algorithm was used to identify the land use types of Landsat-8 remote sensing images [40], and the features in the study area were classified into forest land, arable land, grassland, water and construction land, and the specific process was as follows: firstly, the training samples and validation samples of the different land use types were selected by visual judgment, and then the classification model was trained and the remote sensing images were classified into the land use types, and finally the land use types map (Figure 1C) of the study area was obtained (overall accuracy: 85.7%).

$$NDVI = \frac{(NIR - R)}{(NIR + R)} \quad (2)$$

where NDVI is the normalised vegetation index, *NIR* is the near-infrared band (band 5) of Landsat-8, and R is the red band (band 4) of Landsat-8.

A total of seven feature variables (i.e., geological type, geomorphic type, elevation, slope, TWI, NDVI, and land use type) were extracted from the data sources. All the feature variables were converted to raster data with a resolution of 30 × 30 m and the same number

of rows and columns. Data processing was implemented in ENVI v5.3 (Harris Geospatial Solutions Inc., Broomfield, MA, USA) and ArcGIS v10.2 (Environment System Research Institute Inc., Redlands, CA, USA).

2.4. Training Sample Selection

Extracting a training sample set from the source soil map is a key step in digital soil mapping. The training sample set can be used to identify the soil–environment relationship and build prediction models for updating the soil map or predictive mapping of other factors [5]. The quality of training samples affects the accuracy of the updated soil map [41–43]. To select training samples, the SPM information of all soil survey points was extracted from the data of the Second National Soil Survey. Due to limited techniques at the time of the survey, there may be some sample points with inaccurate information. Therefore, the sample points with incomplete or vaguely defined information on SPM were excluded, such as those with incorrectly documented information, incorrect coordinates, and changes in land use types, and to ensure that the final number of training samples obtained for the different types is similar to the proportion of the area of the SPM types in the historical SPM-type map. A total of 1303 training samples were obtained. The numbers of training samples for RA, QRC, RSWM, GWM, LWM, and MWM were 108, 420, 264, 323, 90, and 98, respectively (Figure 1A).

2.5. Machine Learning Algorithms for Parent Material Classification

RF is a powerful and flexible integrated machine-learning algorithm. It combines the predictions of multiple decision trees by voting to improve the accuracy and stability of the model [30]. The core idea of RF is to randomly select training samples and feature variables for classification. RF uses each decision tree voting to produce results (Equation (3)):

$$\mathrm{A(a)} = B\ \arg_x \max \sum_{y-1}^{C} d_y^x(a) \tag{3}$$

where A(a) is the model based on the RF extractive algorithm; $Barg_x$ is the x-labeling of the extracted class; C is the number of voting decision trees in the forest of the RF extractive algorithm; and d_y is the y-th voting decision tree in the forest of the RF extractive algorithm.

SVM is a machine learning algorithm based on statistical learning theory to implement the structural risk minimisation criterion. It is characterised by small training samples, high noise immunity, and support for high dimensional data [32]. This method also has strong stability and fast classification speed. However, it needs a large number of suitable feature variables to ensure classification accuracy, which limits its application. SVM is proposed as a classifier, which has an outstanding generalisation ability by introducing a kernel function to transform a nonlinear problem into a linear problem. In this study, the radial basis function was used for SVM (Equation (4)):

$$\exp(-\frac{1}{2\sigma^2}||x - x_i||_2) \tag{4}$$

where the inner product kernel K (x, x_i) with width σ^2 is the same for all kernels. Radial basis functions can transform a nonlinear problem into a linear one, improving the accuracy of the classifier and obtaining more accurate classification results [44].

MLC is a nonlinear classification algorithm based on the Bayesian verdict criterion [33]. It assumes that the training samples obey the normal distribution. After training, it calculates the probability of the class corresponding to the grid cell to be classified and then performs the classification. The implementation of MLC is convenient and fast. When combined with the Bayesian theory, this method can effectively classify the object. The discriminant formula of MLC is as follows (Equation (5)):

$$g_i(x) = \ln[p(w_i)] - \frac{1}{2}(x - u_i)^T \sum_i^{-1}(x - u_i) \quad (5)$$

where $P(w_i)$ is the prior probability of the class; w_i is the conditional probability of observing from the class to the raster cell x; i is the number of features; $g_i(x)$ is the likelihood of belonging to the class w_i in x; u_i is the mean vector of class i; and $\sum_i$ is the variance-covariance matrix for class i.

All machine learning algorithms were run with code written in Python v3.7.3 (https://www.python.org/, accessed on 21 June 2023) and implemented by calling machine learning modules in the Sklearn library [45]. During the training process, the training sample set was divided into a training set and a test set (7:3). The optimal parameters of each training model were obtained by grid search. The main parameters to be adjusted for the RF algorithm were *n_estimators* and *max_features*; the main parameters to be adjusted for the SVM algorithm were *class_weight* and *max_iter*; and the main parameters to be adjusted for the MLC algorithm were *probability_threshold* (See Appendix A Table A1 for specific parameter settings).

2.6. Accuracy and Uncertainty Assessment of Parent Material Classification

A validation sample set was obtained by stratification to assess the accuracy of SPM classification results. First, a map of soil sampling points was designed based on the historical maps of soil type, land use type, geological type, elevation distribution, and geomorphic type. (i) There were sampling points distributed in each soil type, land use type, geological type, elevation interval (100 m), and geomorphic type. (ii) The sampling points were designed in equal proportions based on the area ratios of different soil types, land use types, geological types, elevation intervals, and geomorphologic types. (iii) The sampling points were generally located at the centre of patches of various SPM types, with no points on patch boundaries. Then, SPM types were surveyed in the field. At each sampling point, the soil profile was dug into the SPM layer using a soil extractor or shovel and used by a soil scientist to determine the SPM type. At the same time, information on elevation, slope, land use type, soil type, etc., of the survey sample site location was calibrated. A total of 586 validation sampling points were surveyed. The numbers of validation samples for RA, QRC, LWM, RSWM, GWM, and MWM were 84, 181, 107, 137, 49, and 28, respectively (Figure 1A). In Shanggao County, there are field validation sample points in most areas, of which the southern mountainous area is larger, and the sample points are more densely distributed, and the central and northern parts of the county are mostly plains and hills, and the distribution of sample points is more dispersed.

The validation sample set was used to assess the accuracy of the conventional SPM-type map derived from historical soil maps and the digital SPM-type maps obtained by different classification methods. Confusion matrices were constructed based on the 586 validation samples. The overall accuracy (OA), precision (P), recall (R), F1 score (F1), and Kappa coefficient were calculated based on Equations (6)–(10) [46]:

$$Overall\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Rrcall = \frac{TP}{TP + FN} \quad (8)$$

$$F1score = \frac{2Precision \times Recall}{Precision + Recall} \quad (9)$$

$$Kappa\ coefficient = \frac{P_0 - P_e}{1 - P_e} \quad (10)$$

where TP is the true positive rate, i.e., the number of validation samples correctly identified as a certain SPM type; FP is the false positive rate, i.e., the number of validation samples incorrectly identified as a certain SPM type; TN is the true negative rate, i.e., the number of validation samples correctly identified as other SPM types; FN is the false negative, i.e., the number of validation samples incorrectly identified as other SPM types; F1 score is a weighted average of precision and recall; P_0 is the proportion of correctly classified sampling points; and P_e is the probability of random consensus.

The OA values fall in the range of [0, 1], and values closer to 1 indicate higher overall accuracy of the classification results. The P and R values are in the interval of [0, 1], and values closer to 1 indicate a higher probability of correctly classifying a certain type of matrices and a lower probability of incorrectly identifying other types of matrices. The F1 values, which are the harmonic mean of P and R, are in the range of [0, 1]; values closer to 1 indicate a smaller conflict between P and R, which means better classification results. The Kappa coefficients are between [−1, 1], and values closer to 1 indicate higher consistency of the classification results for all SPM types.

In order to determine the stability of the final soil matrices classification results, the uncertainty of the selected best soil matrices classification results was assessed using the bootstrap algorithm [47]; this algorithm is a robust test without preconditions, which allows the stability of the classification results to be visually assessed by various statistics [48]. The training samples were retrained 100 times by randomly dividing them into a training set and a test set in a ratio of 7:3, and then the field validation samples were used to calculate the mean, standard deviation, maximum and minimum values of the overall accuracy and Kappa coefficients of the obtained classification results of these soil matrices in order to assess the uncertainty of the best classification results. If the standard deviation, the difference between the maximum and minimum values and the difference between the mean value and the best classification results are smaller, the uncertainty of the classification result is lower; otherwise, the uncertainty is higher.

2.7. Evaluation of Feature Importance

SHAP provides a unified explanatory framework for all complex machine-learning models [49]. This method links classical Shapley values from game theory to local explanations to quantify the marginal contribution of each input feature to individual sample predictions. It is expressed by the following (Equation (11):

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \tag{11}$$

where $z' \in \{0,1\}^M$ represents the presence or absence of feature variables; M is the number of feature variables in the model; ϕ_0 is the constant when all inputs are missing; and ϕ_i is the marginal contribution of variable i, i.e., the Shapley value.

The SHAP method starts from individual sample predictions. It can assess not only the global importance but also the local importance of feature variables, providing sufficient details for model interpretation. The feature importance in SPM identification models was calculated using the SHAP package in Python v3.7.3 (https://www.python.org/, accessed on 13 August 2023). A larger importance value indicates that the feature variable contributes more to the classification results.

3. Results

3.1. Accuracy of Classification Models with Different Feature Combination Schemes

SPM classification models were developed using the training sample set based on RF, SVM, and MLC algorithms with different feature combination schemes. The accuracy of SPM identification results based on different methods was compared in Figure 3. MLC showed the lowest performance in identifying SPM types, as indicated by its overall accuracy, Kappa coefficient, and F1 score (no higher than 36%, 0.22, and 0.40, respectively).

The accuracy of MLC-based results did not improve with an increasing number of feature variables, indicating that this algorithm was not applicable to the identification of SPM types in the study area.

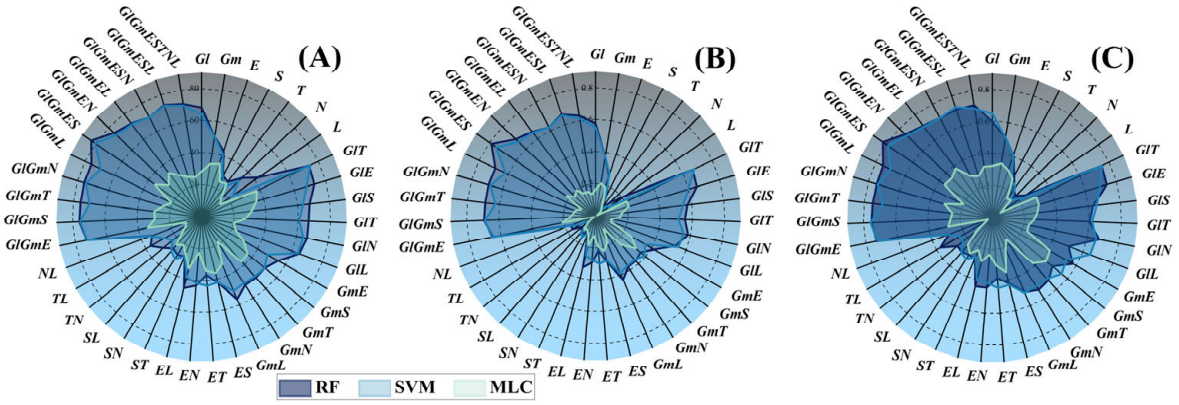

Figure 3. Comparison of the (**A**) overall accuracy, (**B**) Kappa coefficient, and (**C**) F1 score of random forest (RF), support vector machine (SVM), and maximum likelihood classification (MLC) models based on different feature combination schemes for the identification of soil parent maternal types. *Gl*—geological type, *Gm*—geomorphic type, *E*—elevation, *S*—slope, *T*—topographic wetness index, *N*—normalised difference vegetation index, and *L*—land use type. For example, *GlGmES* represents the geological type + geomorphic type + elevation + slope scheme.

The accuracy of the results based on RF and SVM algorithms using different feature combination schemes showed consistent trends. Overall, the RF- and SVM-based results obtained with the *GlGmDS* scheme showed the highest accuracy in terms of overall accuracy (83.11% and 80.20%), Kappa coefficient (0.79 and 0.75), and F1 score (0.82 and 0.79). When using single feature variables to identify SPM types, the highest accuracy was obtained with geological type, followed by geomorphic type and elevation. When using two-feature combinations, the identification accuracy obtained with the schemes, including the geological type or geomorphic type, was higher than that obtained with other schemes. The highest accuracy of RF- and SVM-based results was obtained using the *GlGm* scheme, with overall accuracy of 73.72% and 74.57%, Kappa coefficients of 0.67 and 0.68, and F1 scores of 0.75 and 0.75, respectively.

On the basis of *GlGm*, more feature variables were added to obtain multi-feature combinations. The identification accuracy of SPM types with three-feature combinations was considerably higher than that with single-feature variables and two-feature combinations. Among the three feature combinations, the overall accuracy of the results obtained with the *GlGmE* scheme was the highest. Therefore, more feature variables were continuously selected for integration with *GlGmE*. When using the four-feature combinations, the best results were obtained with the *GlGmES* scheme. Further selection of more feature variables did not improve the identification accuracy and even compromised the accuracy compared with *GlGmES*, which suggests that selecting too many feature variables would lead to information redundancy.

The precision and recall of the results obtained by RF and SVM algorithms with different feature combination schemes are compared in Figure 4. MLC was not considered here because of its low overall accuracy. With respect to individual SPM types, the identification results of RA, QRC, and RSWM based on the GlGmES-RF model showed the highest precision and recall compared with the results obtained using other feature combination schemes. Although several models had higher precision (recall) for LWM, MWM, and GWM, their recall (precision) was lower than that of the GlGmES-RF model. In terms of precision and recall, *GlGmES* was superior to other schemes in identifying various SPM types. This confirmed that *GlGmES* was the optimal feature combination scheme for SPM identification in the study area.

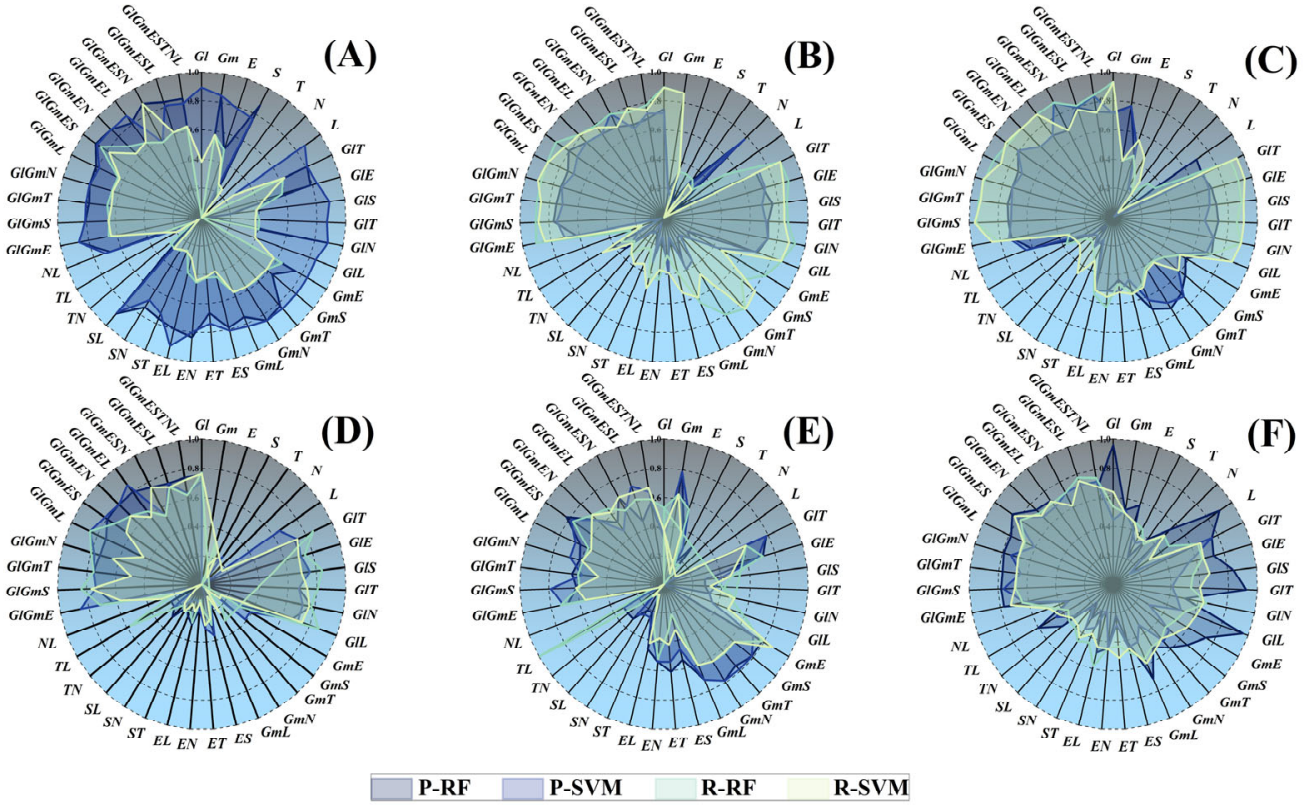

Figure 4. Comparison of the precision and recall of random forest (RF) and support vector machine (SVM) models based on different feature combination schemes. (**A**–**F**) show the identification results of RA, QRC, RSWM, LWM, MWM, and GWM, respectively.

While most models with different feature combinations were able to identify each of the six SPM types, a few models failed to identify all the six SPM types. For example, the Gl-RF model did not perform well in identifying RA at any validation sampling points. Similarly, the Gm-RF and Gm-SVM models failed to identify MWM, whereas the N-RF and N-SVM models were unable to identify RA, MWM, and GWM. Additionally, the GlL-RF and GmL-RF models could not identify RA, MWM, and GWM. All these models used single- or two-feature combination schemes, indicating that only specific SPM types can be identified by integrating a small number of feature variables.

The performance of the three machine learning algorithms in SPM identification was ranked as follows: RF > SVM > MLC. The MLC algorithm produced poor results with low accuracy, whereas the SVM algorithm achieved higher accuracy than RF based on specific feature combination schemes, such as *GlGm*. In other cases, the RF algorithm obtained higher accuracy than the SVM algorithm, and the consistency of RF-based results for various SPM types was higher than that of SVM-based results, indicating that the RF algorithm had better applicability in SPM identification. Specifically, the mean values of overall accuracy, Kappa coefficient, and F1 score for all SPM classification results obtained by RF were 54.84%, 0.43, and 0.53, respectively; the mean values of the three SPM classification results obtained by SVM were 52.28%, 0.41, and 0.52, respectively, and those of MLC were 27.01%, 0.15, and 0.26. The mean values of overall accuracy, Kappa coefficient, and F1 score obtained by GlGmES-RF were 83.11%, 0.79, and 0.82, and all three accuracy ratings were greater than those of the other models. Comparing the precision and recall of the recognition results of each matrix type, the three machine-learning models also present the trend of RF > SVM > MLC.

So, comparing the results of 78 models based on 39 feature combination schemes revealed that *GlGmES* was the optimal feature combination scheme for SPM identification, with the GlGmES-RF model achieving the highest accuracy.

Table 1 shows the training samples of GlGmES-RF and GlGmES-SVM were re-divided randomly into training and testing sets in the ratio of 7:3, and the mean, standard deviation, maxima and minima of overall accuracy and Kappa coefficients of the SPM classification results were obtained after 100 times of training. From Table 1, we can learn that the mean values of the overall accuracy of the classification results obtained by retraining GlGmES-RF and GlGmES-SVM 100 times are 82.97% and 80.02%, which are 0.13% and 0.18% different from the original overall accuracy of the classification results obtained, while the difference between the mean values of Kappa is 0.01. The standard deviation, maximum and minimum values of overall accuracy for the classification results obtained by retraining GlGmES-RF 100 times are 0.047%, 83.67% and 82.14%, and the standard deviation, maximum and minimum values of Kappa are 0.019, 0.83, and 0.77, and the difference between the standard deviation and the maximum and minimum values are all smaller than the values in GlGmES-SVM. This indicates that the SPM classification results obtained by GlGmES-RF show a lower level of uncertainty, and their stability is better than that of the GlGmES-RF matrices classification results. It also indicates that the SPM types mapping and validation utilizing these training and validation samples have a certain level of representativeness and stability.

Table 1. Comparison of uncertainty assessment values of soil parent material (SPM) classification results from GlGmES-RF and GlGmES-SVM.

Assessed Value	GlGmES-RF		GlGmES-SVM	
	OA/%	Kappa	OA/%	Kappa
Mean value	82.97	0.78	80.02	0.74
Standard deviation	0.047	0.019	0.053	0.023
Maximum value	83.67	0.83	81.13	0.71
Minimum value	82.14	0.77	79.12	0.78

3.2. Comparison of Conventional and Digital Maps of Parent Material Types

Given the poor accuracy of the MLC algorithm in identifying SPM types, only the results based on RF and SVM algorithms with the *GlGmES* scheme were compared with the conventional SPM-type map. Confusion matrices constructed based on the 586 validation samples are shown in Figure 5. The overall accuracy of the conventional map was 57.00%, and its Kappa coefficient was 0.47. The GlGmES-RF model obtained markedly improved results in terms of overall accuracy (by 26%) and Kappa coefficient (by 0.31). The boundaries of QRC, RSWM, MWM, and LWM were indistinct in the conventional map, with large areas of confusion (Figure 5A). The *GlGmES* scheme supported the identification of SPM types well. The GlGmES-RF model did not perform as well as the GlGmES-SVM model in determining the fuzzy areas of QRC and LWM. In all other areas, the GlGmES-RF model showed better performance in SPM identification compared with the GlGmES-SVM model (Figure 5B,C).

As the conventional map (Figure 1B) was limited by mapping scale and manual mapping technique, it could not display small patches or provide detailed information on SPM types. Compared with the conventional map, the digital maps produced based on the GlGmES-RF (Figure 6A) and GlGmES-SVM (Figure 6B) models had a finer resolution. Both digital maps showed a reasonable number and distribution of patches, with clear boundaries between patches of different SPM types. Comparing the two digital maps revealed indistinct differences in the distribution of SPM types. Nevertheless, in the map based on GlGmES-SVM, more patches were finely fragmented in some areas, and this phenomenon was not prominent in the map based on GlGmES-RF. Moreover, the accuracy of the GlGmES-RF model was slightly better than that of the GlGmES-SVM model. Therefore, GlGmES-RF was selected as the optimal model for the identification of SPM types in the study area. The key parameters of the GlGmES-RF model were as follows: *max_features* = 4 and *n_estimators* = 500.

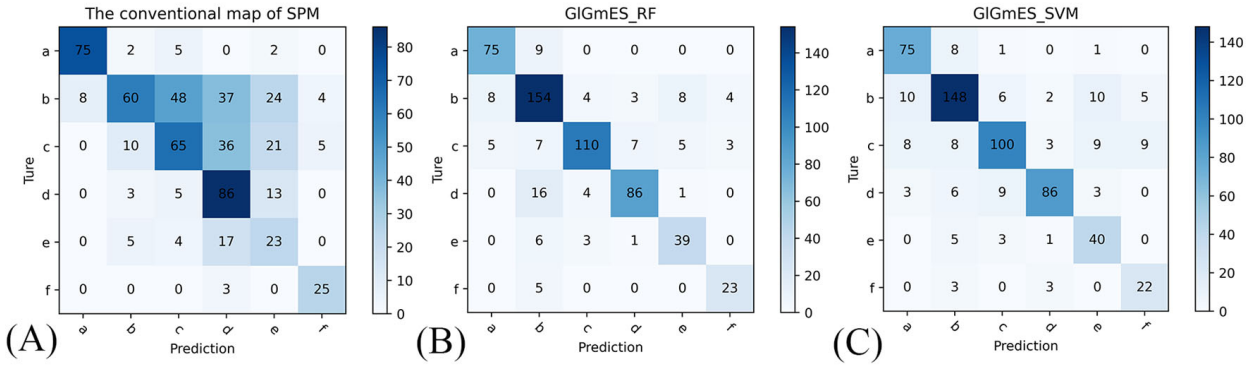

Figure 5. Comparison of confusion matrices based on the validation sample set for (**A**) the conventional map of soil parent material types and the identification results of (**B**) GlGmES-RF and (**C**) GlGmES-SVM models. a–f are RA, QRC, RSWM, LWM, MWM, and GWM, respectively.

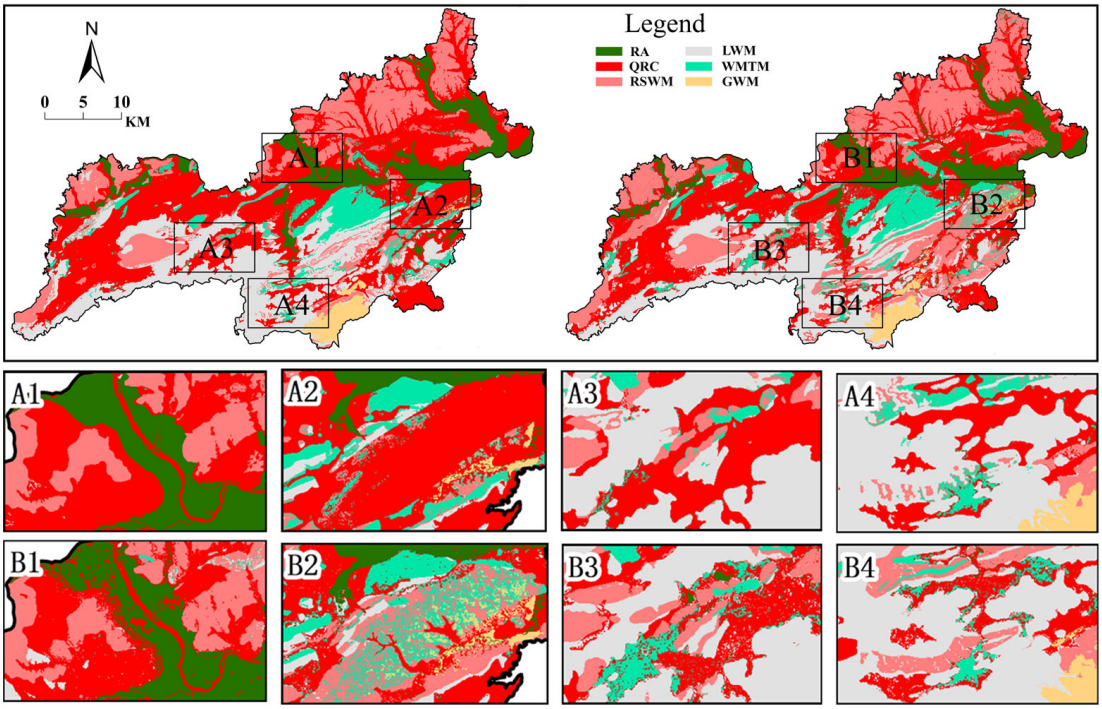

Figure 6. Comparison of the distribution of soil parent material types in two digital maps based on (**A**) GlGmES-RF and (**B**) GlGmES-SVM models.

Table 2 compares the area of different SPM types between the conventional map based on a field survey and two digital maps based on machine learning models. Based on the results obtained by the optimal GlGmES-RF model, the most widely distributed SPM type was QRC, followed by RSWM, and the least distributed SPM type was GWM. Specifically, QRC was found from west to east across the central part of the study area, and RSWM mainly occurred in the northern part. LWM was primarily located in the southern and central parts of the study area, with MWM only found in the central part. RA was chiefly distributed near the banks of the Jinjiang River and its tributaries, whereas GWM only existed in the highest peak area of southern Shanggao County. The area ratios of various SPM types identified by the GlGmES-RF and GlGmES-SVM models did not differ substantially. However, with regard to QRC and LWM, there were large discrepancies in the conventional map compared with the two digital maps. Compared to GlGmES-RF, the QRC changed from 17.85% to 29.10% and 34.57%, and LWM changed from 36.84%

to 17.68% and 18.57%. This suggests there were large confusion areas of QRC and LWM distribution in the conventional map, consistent with the pattern observed in Figure 5.

Table 2. Areas and ratios of different soil parent material (SPM) types in the study area based on digital and conventional mapping.

SPM Type	GlGmES-RF		GlGmES-SVM		Conventional Map	
	Area/hm^2	Ratio/%	Area/hm^2	Ratio/%	Area/hm^2	Ratio/%
River alluvium	12,608.64	9.35	11,485.53	8.52	9833.27	7.29
Quaternary red clay	39,238.11	29.10	46,607.40	34.57	24,062.67	17.85
Red sandstone weathered material	36,601.29	27.15	36,410.31	27.01	30,928.41	22.94
Limestone weathered material	23,836.50	17.68	25,039.08	18.57	49,662.79	36.83
Weathered mudstone material	18,467.91	13.70	12,087.27	8.97	17,192.53	12.75
Granite weathered material	4074.21	3.02	3197.07	2.37	3147.00	2.33

3.3. Feature Importance for Parent Material Identification

The distribution of feature importance for SPM identification was obtained using SHAP (Figure 7). In the GlGmES-RF model with the optimal feature combination, the relative importance scores of feature variables for the overall performance of SPM identification were ordered as geological type (100%) > geomorphic type (78%) > elevation (67%) > slope (54%). However, these feature variables did not contribute greatly to the identification of all SPM types. RSWM, LWM, MWM, and GWM identification were prominently influenced by geological type. Elevation mainly contributed to GWM identification, and geomorphic type played a leading role in RA identification. Although QRC identification was strongly influenced by the geological type, elevation and slope also contributed to this identification process.

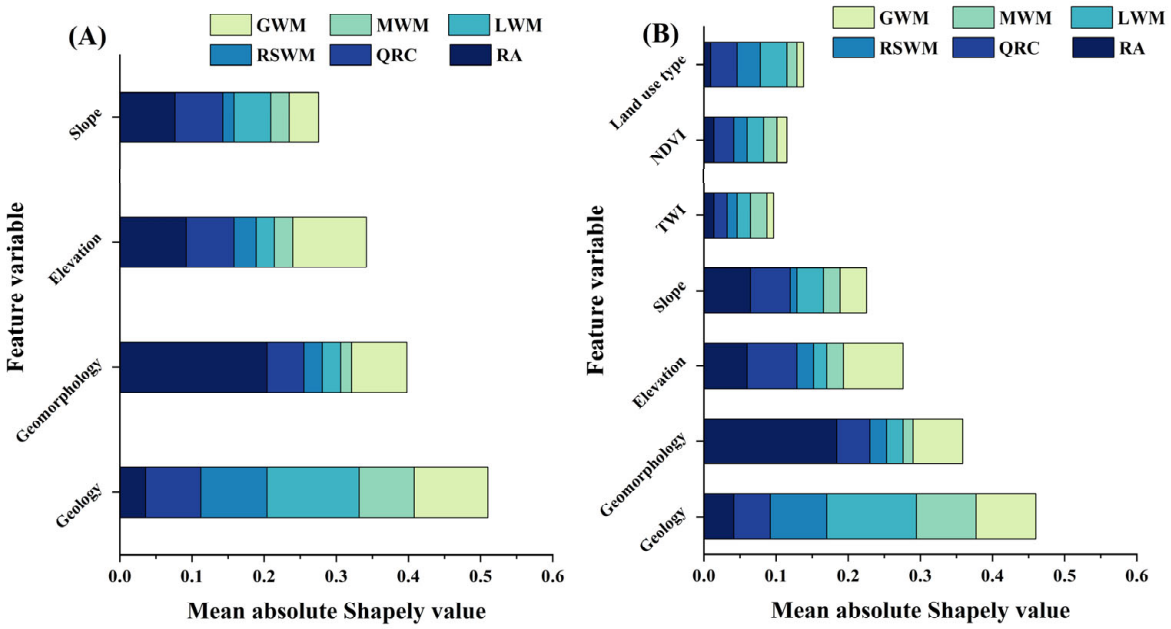

Figure 7. Distribution of feature importance in (**A**) GlGmES-RF and (**B**) GlGmESTNL-RF models for the identification of soil parent material types based on Shapley additive explanations. NDVI—normalised difference vegetation index; TWI—topographic wetness index.

In the GlGmESTNL-RF model with the full feature combination, the relative importance scores of feature variables to the overall performance of SPM identification were ordered as geological type (100%) > geomorphic type (74%) > elevation (60%) > slope (49%) > land use type (30%) > NDVI (25%) > TWI (21%; Figure 7B). The first four feature

variables showed consistent contribution to SPM identification as observed in the GlGmES-RF model. The relative importance scores of land use type, NDVI, and TWI were more than 30% each, indicating their non-significant role in identifying various SPM types in the study area.

4. Discussion

4.1. Contribution of Feature Variables to Soil Parent Material Classification

In this study, the selected feature variables made different contributions to county-level SPM identification in a hilly area based on machine learning algorithms, mainly because of the differences in the formation process of various SPM types [50]. RSWM, LWM, MWM, and GWM are residual parent materials derived from in situ rock weathering without transport, which means that rock type determines these SPM types [31]. Consequently, these SPM types are tied closely to the lithology information contained in geological type [51]. Additionally, GWM mainly occurs in mountainous areas with an elevation above 800 m, where elevation has a prominent effect on SPM identification. RA is principally formed in alluvial plains and larger gully basins, with the geomorphic type having a strong influence on SPM formation [52]. QRC consists of ancient residual and alluvial deposits formed in the Quaternary–Pliocene period. The distribution of residual material can be directly indicated by geological type, whereas alluvial material is transported from other areas and usually distributed in low hills and gentle slopes [53,54]. Therefore, QRC identification is in closely association with geological type, elevation, and slope.

Among the seven feature variables extracted from different data sources, geological type, geomorphic type, elevation, and slope could well explain the distribution patterns of SPM types in the study area. The major rock type in Shanggao County is red sandy shale, which is responsible for RSWM formation. The landform here is dominated by hills, and red clay is likely to accumulate on gentle slopes, leading to QRC formation. Due to the influence of water flow, the Jinjiang River and its tributaries form alluvial plains, which support RA development. High mountains with an elevation of more than 800 m are only found in the southern part of Shanggao County, where GWM results from the dominant rock type—granite. Deciphering the relationship between these feature variables and SPM types is essential for the sustainable exploitation of soil resources. In the case of Shanggao County, the distribution area of RA provides favourable soil water and fertility conditions for planting major crops such as rice, whereas other SPM types, including RSWM and GWM, are mainly distributed in mountainous and hilly areas suitable for tree planting.

There are possible differences in TWI, NDVI, and land use type with various SPM types. However, TWI only changes at large scales, whereas in small regions (especially hilly areas), the difference in topography is more evident. While the present study only used a single-phase remote sensing image to obtain NDVI, multi-phase remote sensing images are often required to link this variable to SPM types [55–57]. Land use types vary in many areas as a result of anthropogenic activities, with SPM having a limited impact. This study was conducted in a county where TWI, NDVI, and land use type did not characterise the differences among various SPM types well. Consequently, the identification accuracy of SPM types did not improve and was even compromised by continuously adding more feature variables to the *GlGmES* scheme. Our findings are in agreement with the results of Mello et al. [6], which used the RF algorithm with remotely sensed and topographic factors (elevation, slope, and TWI) for SPM classification. Among them, elevation was the most important topographic variable [5], whereas TWI contributed the least to SPM classification.

RA is associated with long-term water scouring and sediment transport in rivers. These processes need to be monitored over a long time span. Due to limited data availability, feature variables related to river hydrology were not selected in this study. Additionally, a growing number of studies have used multi-spectral or hyper-spectral remote sensing data for SPM identification [58,59]. This study only used a single-phase remote sensing image for the extraction of NDVI and land use type. Future studies should extract relevant feature variables from multi-source and multi-phase remote sensing images combined with other

data sources [60]. In this study, the feature variables were obtained and screened from topography hydrology and combined with the SPM-type characteristics of the study area, were screened using the SEaTH (Separability and thresholds) algorithm [61], which is more effective in classifying SPM, were retained for the study. Other feature variables, such as the hydrological distance to the nearest stream and aspect, although they have a certain effect on the identification of SPM, were not as effective as TWI and Slope in the context of the present study area. Therefore, these two feature variables were not considered. In addition to the seven feature variables examined in this study, there may be other relevant variables for SPM identification, including land use type and NDVI. It may be helpful to add these feature variables associated with SPM formation to the model in a reasonable manner to improve the accuracy of SPM identification. Further exploration of feature selection can enable more accurate mapping of SPM types.

4.2. Comparison of Different Methods for Mapping of Soil Parent Material Types

In this study, the RF algorithm outperformed SVM and MLC algorithms in the county-level identification of SPM types in a hilly area, consistent with previous studies [58,59]. Since RF-based prediction is the sum of per-tree predictions, this model fully depicts the relationship between each SPM type and feature variables. The RF model takes advantage of this modelling structure to achieve accurate identification of SPM types [62]. Compared with RF, SVM is more sensitive to data scaling [63]. Both geological type and geomorphic types are discontinuous data with data jumps even after normalisation. When the dataset contains an unbalanced number of samples of different types, SVM is likely to predict the SPM type with a higher sample number [64]. Since the number of our training samples for various SPM types was markedly different, SVM cannot reach the same accuracy as RF for SPM identification.

The running time of all the models was considerably different during model training. The training time of RF-based models was between 37–421 s, with a mean of 128 s. The training time of SVM-based models was between 86–1824 s, with a mean of 537 s. Although some models were run for too long, possibly due to other factors of the workstation, SVM-based models generally required longer time for training than RF-based models. If two SPM types are scattered in space, it would be difficult to distinguish between them based on the SVM classification hyperplane [65], and a fragmented map of SPM types would be easily produced during the identification process. This explains why the results from the *GlGmES*-SVM model contained fragmented patches, in contrast with the results from the *GlGmES*-RF model. MLC has limited ability for deep data mining since it can only classify SPM types with high probability based on the information of training samples [66,67]. Thus, the performance of MLC in identifying SPM types is extremely poor. Furthermore, this study attempted to identify SPM types using a convolutional neural network, a deep learning method [68]. The results were even inferior to MLC-based results due to factors such as an insufficient number of training samples.

This study predicted and mapped SPM types in a hilly area based on the principle of digital soil mapping [3]. The relationship between SPM types and environmental feature variables was explored by modelling coupled with spatial and mathematical analyses. This represents a modern technological system that is different from the conventional mapping of SPM types. Feature variables related to the formation of SPM types or indicative of SPM differences were used to construct models based on machine learning algorithms and then generate digital raster maps. The obtained results can characterise spatial variation in the distribution of SPM types in a precise and detailed way. Moreover, the proposed method for high-accuracy county-level SPM mapping inherits the advantages of digital soil mapping, such as its low cost, high recordability, ease of updating, outstanding efficiency, objective consistency, and reliable mapping results [69,70].

4.3. Error Sources and Study Limitations

The accuracy of automatic SPM identification using machine learning algorithms is highly dependent on the quality and resolution of the input data, including geological maps, geomorphic maps, DEMs, and remote sensing imageries. Errors or uncertainties in these datasets can propagate into the SPM identification process, thus affecting the accuracy of the outcomes. Using alternatives of data sources with higher accuracy or lower error is a way to improve the accuracy of the mapping; for example, DEMs from different sources are not equally accurate in different areas. The DEM data used in this study was ASTER GDEM, and future studies may consider using AW3D30, which has a higher accuracy in hilly areas [71] and may improve the accuracy of the results of SPM-type identification.

Selecting representative sampling points has always been a major difficulty in soil mapping [72]. This study collected 586 field validation samples based on stratified random sampling. Despite considering each soil type, land use type, geological type, elevation interval, and geomorphic type, the distribution of validation sampling points still cannot fully represent the change of SPM types throughout the whole study area. This would also lead to the deviation of the validation results from the actual results [8]. How to set up a better representative sample distribution map is worth exploring in depth.

This study compared three machine learning algorithms (RF, SVM, and MLC) for SPM identification and found that RF outperformed the other two algorithms. However, the performance of different algorithms would vary depending on the specific characteristics of the study area and data sources. Choosing other algorithms or combining methods may produce inconsistent results. For example, the convolutional neural network (CNN), which has good applications in remote sensing, ecology, and soil [68,73,74], can be considered for SPM identification in the future.

Shanggao County is a typical hilly area in southern China, with both plains and low hills. Its SPM types include impact matrices, residual matrices, and slope matrices. The method described in this study (i.e., *GlGmES*-RF) is applicable to counties with a similar SPM genesis as Shanggao County. However, for other counties, there may be more appropriate feature variables and better classification algorithms. Therefore, more universal and applicable methods for county-level SPM mapping still need to be developed in future research.

5. Conclusions

This study established an accurate county-level method for SPM mapping in hilly areas using machine learning and optimised feature combinations. The *GlGmES*-RF model achieved >83% accuracy in SPM identification and outperformed conventional mapping. Geological type, geomorphic type, elevation, and slope were identified as key explanatory variables related to SPM formation processes.

The RF algorithm proved superior to SVM and MLC for detailed and consistent SPM identification. Optimisation of feature selection prevented information redundancy from declining model performance. The digital SPM maps generated at fine resolutions captured spatial variability in the distribution of SPM types more effectively than the conventional map.

This research contributes a robust framework transferring the principles of digital soil mapping to SPM-type mapping. Practically, the findings are helpful to rational land use planning by elucidating soil–terrain relationships. The digital datasets can also enable more precise agriculture and ecological monitoring.

Future work should integrate multi-phase remote sensing images and in situ soil property data to strengthen multi-variate modelling. Advancing feature engineering feature selection with topographic wetness indices or colour ratios may enhance SPM classification. Applying the proposed method across larger regions could validate its generalizability.

Overall, the machine learning method developed in this study presents an accurate and efficient solution to address SPM classification challenges in hilly areas. As digital soil information becomes increasingly important, this research provides the foundation for

producing standardised large-area SPM maps that support sustainable land management. Continued methodological refinement will realise the full potential of digital soil mapping for precision agriculture and environmental applications.

Author Contributions: Writing—original draft, X.Z.; Writing—review & editing, X.Z., Y.J., W.L., J.G., Q.Z. and H.Z.; Project administration, X.G.; Funding acquisition, X.G. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Key Research and Development Program of China Project (grant number: 2022YFD1900601-4).

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

NDVI—Normalised difference vegetation index; MLC—Maximum likelihood classification; RF—Random forest; SPM—Soil parent material; SVM—Support vector machine; TWI—Topographic wetness index.

Appendix A

Table A1. Soil parent material (SPM) classification model training parameter information.

Feature Combination	RF		SVM		MLC
	n_estimators	*max_features*	*class_weight*	*max_iter*	*probability_threshold*
Gl	565	6	1	1200	0.59
Gm	586	4	1	1200	0.68
E	475	5	1	1200	0.55
S	507	4	1	1200	0.68
T	519	4	1	1200	0.59
N	583	6	1	1200	0.66
L	440	6	1	1200	0.59
GlT	513	6	1	1300	0.70
GlE	551	7	1	1300	0.70
GlS	560	3	1	1300	0.66
GlT	482	4	1	1300	0.62
GlN	477	6	1	1300	0.67
GlL	532	3	1	1300	0.64
GmE	476	3	1	1300	0.63
GmS	474	5	1	1300	0.62
GmT	446	4	1	1300	0.54
GmN	479	3	1	1300	0.55
GmL	563	6	1	1300	0.58
ES	465	6	1	1300	0.53
ET	470	5	1	1300	0.56
EN	401	4	1	1300	0.67
EL	488	7	1	1300	0.76
ST	558	5	1	1300	0.71
SN	595	7	1	1300	0.75
SL	479	7	1	1300	0.65
TN	414	6	1	1300	0.75
TL	595	7	1	1300	0.70
NL	413	5	1	1300	0.72
GlGmE	469	3	1	1400	0.77
GlGmS	586	3	1	1400	0.67
GlGmT	491	4	1	1400	0.66
GlGmN	528	6	1	1400	0.68
GlGmL	556	5	1	1400	0.69
GlGmES	500	4	1	1400	0.70
GlGmEN	422	4	1	1400	0.68
GlGmEL	504	5	1	1400	0.65
GlGmESN	479	3	1	1300	0.70

Table A1. *Cont.*

Feature Combination	RF		SVM		MLC
	n_estimators	*max_features*	*class_weight*	*max_iter*	*probability_threshold*
GlGmESL	488	4	1	1300	0.72
GlGmESTNL	493	5	1	1300	0.75

Gl—geological type, *Gm*—geomorphic type, *E*—elevation, *S*—slope, *T*—topographic wetness index, *N*—normalised difference vegetation index, and *L*—land use type. For example, *GlGmES* represents the geological type + geomorphic type + elevation + slope scheme.

References

1. Jenny, H. *Factors of Soil Formation*; McGraw-Hill: New York, NY, USA, 1941.
2. Richter, J.; Owens, P.R.; Libohova, Z.; Adhikari, K.; Fuentes, B. Mapping parent material as part of a nested approach to soil mapping in the Arkansas River Valley. *CATENA* **2019**, *178*, 100–108. [CrossRef]
3. McBratney, A.; Santos, M.M.; Minasny, B. On digital soil mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]
4. Kirillova, N.P.; Sileova, T.M.; Ulyanova, T.Y.; Rozov, S.Y.; Smirnova, I.E. Digital large-scale soil parent material map of Chashnikovo Training and Experimental Soil Ecology Center, Moscow State University. *Mosc. Univ. Soil Sci. Bull.* **2017**, *72*, 93–99. [CrossRef]
5. Heung, B.; Bulmer, C.E.; Schmidt, M.G. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. *Geoderma* **2014**, *214–215*, 141–154. [CrossRef]
6. Mello, F.A.; Bellinaso, H.; Mello, D.C.; Safanelli, J.L.; Mendes, W.D.S.; Amorim, M.T.; Gomez, A.M.; Poppiel, R.R.; Silvero, N.E.; Gholizadeh, A.; et al. Soil parent material prediction through satellite multispectral analysis on a regional scale at the Western Paulista Plateau, Brazil. *Geoderma Reg.* **2021**, *26*, e00412. [CrossRef]
7. Jang, H.J.; Dobarco, M.R.; Minasny, B.; McBratney, A. Creating a soil parent material map digitally using a combination of interpretation and statistical techniques. *Soil Res.* **2021**, *59*, 684–698. [CrossRef]
8. Zhu, A.X.; Liu, J.; Du, F.; Zhang, S.J.; Qin, C.Z.; Burt, J.; Behrens, T.; Scholten, T. Predictive soil mapping with limited sample data. *Eur. J. Soil Sci.* **2015**, *66*, 535–547. [CrossRef]
9. Zhu, A.-X.; Band, L.E. A Knowledge-Based Approach to Data Integration for Soil Mapping. *Can. J. Remote Sens.* **1994**, *20*, 408–418. [CrossRef]
10. Hengl, T.; De Jesus, J.M.; Heuvelink, G.B.M.; Gonzalez, M.R.; Kilibarda, M.; Blagotić, A.; Shangguan, W.; Wright, M.N.; Geng, X.; Bauer-Marschallinger, B.; et al. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE* **2017**, *12*, e0169748. [CrossRef]
11. Brammer, H.; Hole, F.D.; Campbell, J.B. Soil Landscape Analysis. *Geogr. J.* **1985**. [CrossRef]
12. Li, J.; Chen, Y.; Zheng, L.; Jiao, X.; Sui, Y. Development of Soil Classification and Research Avenue. *Soil Crop* **2014**, *3*, 146–150.
13. Zhu, A.; Yang, L.; Fan, N.; Zeng, C.; Zhang, G. The review and outlook of digital soil mapping. *Prog. Geogr.* **2018**, *37*, 66–78. [CrossRef]
14. Gruber, F.E.; Baruck, J.; Mair, V.; Geitner, C. From geological to soil parent material maps—A random forest-supported analysis of geological map units and topography to support soil survey in South Tyrol. *Geoderma* **2019**, *354*, 113884. [CrossRef]
15. Armstrong, J.E. *Surficial Geology of Vancouver Area, British Columbia*; Geological Survey of Canada: Ottawa, ON, Canada, 1956; paper 55.
16. Armstrong, J.E. *Surficial Geology of New Westminster Map-Area, British Columbia*; Department of Mines and Technical Surveys: Ottawa, ON, Canada, 1957; Paper 57 and Map 16.
17. Li, A.-D.; Guo, P.-T.; Wu, W.; Liu, H.-B. Impacts of terrain attributes and human activities on soil texture class variations in hilly areas, south-west China. *Environ. Monit. Assess.* **2017**, *189*, 281. [CrossRef]
18. Zhang, D.; Zhou, F.; Fang, K.; Davi, N.; Chen, Z.; Wang, F.; Chen, Y. Quantification of soil erosion dynamics in the hilly red soil region of Southeast China based on exposed roots. *CATENA* **2023**, *232*, 107386. [CrossRef]
19. Grimm, R.; Behrens, T.; Märker, M.; Elsenbeer, H. Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. *Geoderma* **2008**, *146*, 102–113. [CrossRef]
20. Gray, J.M.; Bishop, T.F.; Wilford, J.R. Lithology and soil relationships for soil modelling and mapping. *CATENA* **2016**, *147*, 429–440. [CrossRef]
21. Zeferino, L.B.; de Souza, L.F.T.; Amaral, C.H.D.; Filho, E.I.F.; de Oliveira, T.S. Does environmental data increase the accuracy of land use and land cover classification? *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *91*, 102128. [CrossRef]
22. Zhang, L.; Wang, Y.; Niu, M.; Wang, C.; Wang, Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study. *Sci. Rep.* **2020**, *10*, 4406. [CrossRef]
23. Georganos, S.; Grippa, T.; Vanhuysse, S.; Lennert, M.; Shimoni, M.; Wolff, E. Very high resolution object-based land use–land cover urban classification using extreme gradient boosting. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 607–611. [CrossRef]
24. Sheykhmousa, M.; Mahdianpari, M.; Ghanbari, H.; Mohammadimanesh, F.; Ghamisi, P.; Homayouni, S. Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6308–6325. [CrossRef]

25. Hengl, T.; Heuvelink, G.B.M.; Kempen, B.; Leenaars, J.G.B.; Walsh, M.G.; Shepherd, K.D.; Sila, A.; Macmillan, R.A.; De Jesus, J.M.; Tamene, L.; et al. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. *PLoS ONE* **2015**, *10*, e0125814. [CrossRef]
26. Lacoste, M.; Lemercier, B.; Walter, C. Regional mapping of soil parent material by machine learning based on point data. *Geomorphology* **2011**, *133*, 90–99. [CrossRef]
27. Brevik, E.C.; Miller, B.A. The Use of Soil Surveys to Aid in Geologic Mapping with an Emphasis on the Eastern and Midwestern United States. *Soil Horiz.* **2015**, *56*, 1–9. [CrossRef]
28. Miller, B.A.; Burras, C.L. Comparison of Surficial Geology Maps Based on Soil Survey and In Depth Geological Survey. *Soil Horiz.* **2015**, *56*, 1–12. [CrossRef]
29. Oehlke, B.M.; Dolliver, H.A.S. Quaternary Glacial Mapping in Western Wisconsin Using Soil Survey Information. *J. Nat. Resour. Life Sci. Educ.* **2011**, *40*, 73–77. [CrossRef]
30. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
31. Heung, B.; Ho, H.C.; Zhang, J.; Knudby, A.; Bulmer, C.E.; Schmidt, M.G. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma* **2016**, *265*, 62–77. [CrossRef]
32. Adugna, T.; Xu, W.; Fan, J. Comparison of Random Forest and Support Vector Machine Classifiers for Regional Land Cover Mapping Using Coarse Resolution FY-3C Images. *Remote Sens.* **2022**, *14*, 574. [CrossRef]
33. Mahmoud, R.R.; Mahmood, K.A.; Mahmoud, R.M. Maximum likelihood classification for land cover change detection utilizing lands at-8 imagery in Wasit Governorate, I.R.A.Q. *Ecol. Environ. Conserv.* **2019**, *25*, 68–72.
34. Lan, J.J.; Yu, H.Y.; Chen, L.; Ma, H.H.; Zhang, H.Y. Scaling effect of airborne LiDAR DEM in watershed hydrological analysis and simulation. *Surv. Mapp. Bull.* **2020**, *2020*, 40–46. [CrossRef]
35. de Oliveira, M.F.; Ortiz, B.V.; Morata, G.T.; Jiménez, A.-F.; Rolim, G.d.S.; da Silva, R.P. Training Machine Learning Algorithms Using Remote Sensing and Topographic Indices for Corn Yield Prediction. *Remote Sens.* **2022**, *14*, 6171. [CrossRef]
36. Tang, G.; Yang, X.; Zhou, C.; Li, F.; Xiong, L.; Li, S.; Na, J. Nanjing Normal University. 2023. Global Basic Landform Units Datasets (2023), Version 1. Yangtze River Delta Science Data Center, National Earth System Science Data Sharing Infrastructure, National Science & Technology Infrastructure of China. Available online: http://geodata.nnu.edu.cn/ (accessed on 21 June 2023).
37. Walker, D.A. Hierarchical subdivision of Arctic tundra based on vegetation response to climate, parent material and topography. *Glob. Chang. Biol.* **2000**, *6*, 19–34. [CrossRef]
38. Zhang, L.; Gong, Z.; Wang, Q.; Jin, D.; Wang, X. Wetland mapping of Yellow River Delta wetlands based on multi-feature optimization of Sentinel-2 images. *J. Remote Sens.* **2019**, *23*, 313–326. [CrossRef]
39. Liu, Y.; Li, Z.; Chen, Y.; Li, Y.; Li, H.; Xia, Q.; Kayumba, P.M. Evaluation of consistency among three NDVI products applied to High Mountain Asia in 2000–2015. *Remote Sens. Environ.* **2022**, *269*, 112821. [CrossRef]
40. Peng, X.; He, G.; She, W.; Zhang, X.; Wang, G.; Yin, R.; Long, T. A Comparison of Random Forest Algorithm-Based Forest Extraction with GF-1 WFV, Landsat 8 and Sentinel-2 Images. *Remote Sens.* **2022**, *14*, 5296. [CrossRef]
41. Qi, F.; Zhu, A.-X. Knowledge discovery from soil maps using inductive learning. *Int. J. Geogr. Inf. Sci.* **2003**, *17*, 771–795. [CrossRef]
42. Qi, F. Knowledge Discovery from Area-Class Resource Maps: Data Preprocessing for Noise Reduction. *Trans. GIS* **2004**, *8*, 297–308. [CrossRef]
43. Grinand, C.; Arrouays, D.; Laroche, B.; Martin, M.P. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. *Geoderma* **2008**, *143*, 180–190. [CrossRef]
44. Basheer, S.; Wang, X.; Farooque, A.A.; Nawaz, R.A.; Liu, K.; Adekanmbi, T.; Liu, S. Comparison of Land Use Land Cover Classifiers Using Different Satellite Imagery and Machine Learning Techniques. *Remote Sens.* **2022**, *14*, 4978. [CrossRef]
45. Atienza, R. *Advanced Deep Learning with Keras: Apply Deep Learning Techniques, Autoencoders, GANs, Variational Autoencoders, Deep Reinforcement Learning, Policy Gradients, and More*; Packt Publishing Ltd.: Birmingham, UK, 2018.
46. Congalton, R.G.; Green, K. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, 3rd ed.; Taylor & Francis Group: Boca Raton, FL, USA, 2019.
47. Zhou, T.; Geng, Y.; Chen, J.; Pan, J.; Haase, D.; Lausch, A. 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.* **2020**, *729*, 138244. [CrossRef]
48. Brus, D.; Kempen, B.; Heuvelink, G. Sampling for validation of digital soil maps. *Eur. J. Soil Sci.* **2011**, *62*, 394–407. [CrossRef]
49. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *arXiv* **2017**, arXiv:1705.07874.
50. Halim, M.K.; Ahmad, A.; Rahman, M.Z.; Amin, Z.M.; Khanan, M.F.; Musliman, I.; Kadir, W.H.W.; Jamal, M.H.; Maimunah, D.S.; Wahab, A.K.; et al. Land use/land cover mapping for conservation of UNESCO Global Geopark using object and pixel-based approaches. *IOP Conf. Ser. Earth Environ. Sci.* **2018**, *169*, 012075. [CrossRef]
51. Tazikeh, H.; Khormali, F.; Amini, A.; Motlagh, M.B.; Ayoubi, S. Soil-parent material relationship in a mountainous arid area of Kopet Dagh basin, North East Iran. *CATENA* **2017**, *152*, 252–267. [CrossRef]
52. Da Silva, M.S.; Guimarães, J.T.; Filho, P.W.S.; Júnior, W.N.; Sahoo, P.K.; Da Costa, F.R.; Júnior, R.O.S.; Rodrigues, T.M.; Da Costa, M.F. Morphology and morphometry of upland lakes over lateritic crust, Serra dos Carajás, southeastern Amazon region. *An. Da Acad. Bras. De Ciências* **2018**, *90*, 1309–1325. [CrossRef]

53. Zhu, X.; Liang, Y.; Qu, L.; Cao, L.; Tian, Z.; Liu, T.; Li, M. Characteristics of runoff and sediment yield for two typical erodible soils in southern China. *Int. J. Sediment Res.* **2022**, *37*, 653–661. [CrossRef]
54. Wu, H.; Song, X.; Liu, F.; Li, D.; Zhang, G. Geophysical and geochemical characterization reveals topography controls on critical zone structure in a low hilly region. *Earth Surf. Process. Landf.* **2022**, *47*, 2796–2810. [CrossRef]
55. Dematte, J.A.M.; Huete, A.R.; Ferreira, L.G., Jr.; Nanni, M.R.; Alves, M.C.; Fiorio, P.R. Methodology for Bare Soil Detection and Discrimination by Landsat TM Image. *Open Remote Sens. J.* **2009**, *2*, 24–35. [CrossRef]
56. Izawa, M.R.; Cloutis, E.A.; Rhind, T.; Mertzman, S.A.; Applin, D.M.; Stromberg, J.M.; Sherman, D.M. Spectral reflectance properties of magnetites: Implications for remote sensing. *Icarus* **2019**, *319*, 525–539. [CrossRef]
57. Silva, L.S.; Júnior, J.M.; Barrón, V.; Gomes, R.P.; Teixeira, D.D.B.; Siqueira, D.S.; Vasconcelos, V. Spatial variability of iron oxides in soils from Brazilian sandstone and basalt. *CATENA* **2019**, *185*, 104258. [CrossRef]
58. Bonfatti, B.R.; Demattê, J.A.; Marques, K.P.; Poppiel, R.R.; Rizzo, R.; Mendes, W.d.S.; Silvero, N.E.; Safanelli, J.L. Digital mapping of soil parent material in a heterogeneous tropical area. *Geomorphology* **2020**, *367*, 107305. [CrossRef]
59. Mancini, M.; Weindorf, D.C.; Chakraborty, S.; Silva, S.H.G.; dos Santos Teixeira, A.F.; Guilherme, L.R.G.; Curi, N. Tracing tropical soil parent material analysis via portable X-ray fluorescence (pXRF) spectrometry in Brazilian Cerrado. *Geoderma* **2019**, *337*, 718–728. [CrossRef]
60. Richer-De-Forges, A.C.; Chen, Q.; Baghdadi, N.; Chen, S.; Gomez, C.; Jacquemoud, S.; Martelet, G.; Mulder, V.L.; Urbina-Salazar, D.; Vaudour, E.; et al. Remote Sensing Data for Digital Soil Mapping in French Research—A Review. *Remote Sens.* **2023**, *15*, 3070. [CrossRef]
61. Nussbaum, S.; Niemeyer, I.; Canty, M. SEATH—A New Tool for Automated Feature Extraction in the Context of Object-Based Image Analysis. In Proceedings of the 1st International Conference on Object-Based Image Analysis (OBIA 2006), Salzburg, Austria, 4–5 July 2006.
62. Qi, Y. Random forest for bioinformatics. *Ensemble Mach. Learn. Methods Appl.* **2012**, *4*, 307–323. [CrossRef]
63. Xie, G.; Niculescu, S. Mapping and Monitoring of Land Cover/Land Use (LCLU) Changes in the Crozon Peninsula (Brittany, France) from 2007 to 2018 by Machine Learning Algorithms (Support Vector Machine, Random Forest, and Convolutional Neural Network) and by Post-classification Comparison (PCC). *Remote Sens.* **2021**, *13*, 3899. [CrossRef]
64. Sothe, C.; De Almeida, C.M.; Schimalski, M.B.; La Rosa, L.E.C.; Castro, J.D.B.; Feitosa, R.Q.; Dalponte, M.; Lima, C.L.; Liesenberg, V.; Miyoshi, G.T.; et al. Comparative performance of convolutional neural network, weighted and conventional support vector machine and random forest for classifying tree species using hyperspectral and photogrammetric data. *GIScience Remote Sens.* **2020**, *57*, 369–394. [CrossRef]
65. Sevani, N.; Azizah, K.; Jatmiko, W. A Feature-based Transfer Learning to Improve the Image Classification with Support Vector Machine. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*, 291–301. [CrossRef]
66. Ahmad, A.; Kalsom, U.; Mohd, O.; Mawardy, M.; Sakidin, H.; Wahid, A.; Firdaus, S. Comparative Analysis of Support Vector Machine, Maximum Likelihood and Neural Network Classification on Multispectral Remote Sensing Data. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 529–537. [CrossRef]
67. Verbovšek, T.; Popit, T. GIS-assisted classification of litho-geomorphological units using Maximum Likelihood Classification, Vipava Valley, SW Slovenia. *Landslides* **2018**, *15*, 1415–1424. [CrossRef]
68. Zhong, L.; Guo, X.; Xu, Z.; Ding, M. Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks. *Geoderma* **2021**, *402*, 115366. [CrossRef]
69. Cahyana, D.; Sulaeman, Y.; Barus, B.; Darmawan; Mulyanto, B. Improving digital soil mapping in Bogor, Indonesia using parent material information. *Geoderma Reg.* **2023**, *33*, e00627. [CrossRef]
70. Lagacherie, P.; McBratney, A. Spatial soil information systems and spatial soil inference systems: Perspectives for digital soil mapping. *Dev. Soil Sci.* **2007**, *31*, 3–24.
71. Nadi, S.; Shojaei, D.; Ghiasi, Y. Accuracy Assessment of DEMs in Different Topographic Complexity Based on an Optimum Number of GCP Formulation and Error Propagation Analysis. *J. Surv. Eng.* **2020**, *146*, 04019019. [CrossRef]
72. Liu, J.; Zhu, A.; Zhang, S.; Qin, C. Large-scaled soil attribute mapping method based on individual representativeness of sample sites. *Acta Pedologica Sinica* **2013**, *50*, 12–20.
73. Zhong, L.; Guo, X.; Guo, J.X.; Xu, Z.; Zhu, Q.; Ding, M. Hyperspectral estimation of organic matter in red soil using different convolutional neural network models. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 203–212. [CrossRef]
74. De Bem, P.P.; de Carvalho Junior, O.A.; Fontes Guimarães, R.; Trancoso Gomes, R.A. Change Detection of Deforestation in the Brazilian Amazon Using Landsat Data and Convolutional Neural Networks. *Remote Sens.* **2020**, *12*, 901. [CrossRef]