remote sensing

MDPI

Article

Soil Classification Mapping Using a Combination of Semi-Supervised Classification and Stacking Learning (SSC-SL)

Fubin Zhu, Changda Zhu, Wenhao Lu, Zihan Fang, Zhaofu Li and Jianjun Pan *

College of Resources and Environmental Sciences, Nanjing Agricultural University, No. 1 Weigang, Xuanwu District, Nanjing 210095, China
* Correspondence: jpan@njau.edu.cn

Abstract: In digital soil mapping, machine learning models have been widely applied. However, the accuracy of machine learning models can be limited by the use of a single model and a small number of soil samples. This study introduces a novel method, semi-supervised classification combined with stacking learning (SSC-SL), to enhance soil classification mapping in hilly and low-mountain areas of Northern Jurong City, Jiangsu Province, China. This study incorporated Gaofen-2 (GF-2) remote sensing imagery along with its associated remote sensing indices, the ALOS Digital Elevation Model (DEM) and their derived topographic factors, and soil parent material data in its modelling process. We first used three base learners, Ranger, Rpart, and XGBoost, to construct the SL model. In addition, we employed the fuzzy c-means clustering algorithm (FCM) to construct a clustering map. To fully leverage the information from a multitude of environmental variables, understand the distribution of data, and enhance the effectiveness of the classification, we selected unlabelled samples near the boundaries of the patches on the clustering map. The SSC-SL model demonstrated superior stability and performance, with optimal accuracy at a 0.9 confidence level, achieving an overall accuracy of 0.77 and a kappa coefficient of 0.73. These metrics exceeded those of the highest performing base learner (Ranger model) by 10.4% and 12.3%, respectively, and they outperformed the least effective base learner (Rpart model) by 27.3% and 32.9%. It notably improves the spatial distribution accuracy of soil types. Key environmental variables influencing soil type distribution include soil parent material (SPM), land use (LU), the multi-resolution valley bottom flatness index (MRVBF), and Elevation (Ele). In conclusion, the SSC-SL model offers a novel and effective approach for enhancing the predictive accuracy of soil classification mapping.

Keywords: SSC-SL; stacking learning; semi-supervised classification; FCM; digital soil mapping; GF-2

Citation: Zhu, F.; Zhu, C.; Lu, W.; Fang, Z.; Li, Z.; Pan, J. Soil Classification Mapping Using a Combination of Semi-Supervised Classification and Stacking Learning (SSC-SL). *Remote Sens.* **2024**, *16*, 405. https://doi.org/10.3390/rs16020405

Academic Editor: Jeroen Meersmans

Received: 3 December 2023
Revised: 14 January 2024
Accepted: 18 January 2024
Published: 20 January 2024

1. Introduction

Soil is a loose layer of material on the Earth's surface and is an important part of the Earth's ecosystem [1]. It provides a foundation for plant growth and a habitat for animals, while also storing water and nutrients. Soils also play a key role in climate regulation and food security [2]. It is therefore essential to understand the distribution of soil types and to produce accurate soil type maps. Soil maps are useful for making optimal land use decisions [3] and analysing land suitability [4].

Recently, machine learning, as a method often applied to regression and classification tasks in various scientific fields, has become more widely used in digital soil mapping and has gradually become one of the most mainstream methods in digital soil mapping. The machine learning applications in soil attribute mapping are mainly used for determining organic carbon [5], pH [6], soil texture [7], soil temperature [8], and soil salinity [9]. Similarly, many studies have used machine learning to predict soil type maps. For example, Najmeh Asgari et al. [10] produced soil maps of the Jouneqan district in Chaharmahal and Bakhtiary Province, Iran. Teng et al. [11] updated the Australian soil classification map. These studies used various machine learning methods, including quantile regression forests, random

forests, support vector machines, and multinomial logistic regression. These methods achieve the rapid prediction of soil properties and types by fitting non-linear relationships between environmental variables and soil properties and types and have proven to be a well-established tool.

However, the current problem is that while a single machine learning method can predict soil types, sometimes it does not perform well in terms of prediction accuracy. The solution to this problem is to have the ability to combine multiple machine learning methods to improve the prediction accuracy in soil classification [12,13]. Stacking learning (SL) is a branch of ensemble learning that combines multiple base learners into a single advanced learner. Typically, stacking learning is robust and adaptive and can achieve better results than a single base learner [14]. In addition, in digital soil mapping research, especially in soil type mapping, the soil sampling point dataset is, in most cases, a small sample dataset. A small sample point dataset sometimes cannot capture the environmental variables corresponding to each soil type very well. Therefore, the prediction results are not a good representation of the real soil distribution when using machine learning methods for predictive mapping. It is worth noting that in many studies of soil type mapping, a small number of actual sampling point data are often used for training and prediction, while large numbers of unlabelled environmental variable data are not fully utilised [15,16]. In contrast, semi-supervised classification (SSC) methods are based on learning from labelled data and predictive classification from unlabelled data, resulting in some unlabelled data for training [17]. Semi-supervised classification makes full use of the existing data to improve the classification accuracy of the model when data resources are limited [18]. Therefore, semi-supervised classification can be used to achieve higher soil mapping accuracy in digital soil mapping.

Currently, although some studies have proposed the application of stacking learning in soil attribute prediction [19] and semi-supervised classification in soil type prediction [20], we have not seen any application of stacking learning for soil type prediction or for the research application of combining stacking learning and semi-supervised classification for soil type prediction. Furthermore, in the selection of unlabelled sample points in semi-supervised classification and considering the need to ensure a more reasonable delineation of soil type patch boundaries, we employ the fuzzy c-means clustering algorithm (FCM) to obtain clustering results and add the positions of unlabelled sample points near the boundaries of the clustering results. This is because unlabelled samples closer to the centre of the patches have more similar environmental variables to labelled samples, and selecting unlabelled samples closer to the centre of the patches can make the model less stable. The aims of this study are to combine multiple base learners and make the best use of a large volume of environmental variable data by combining stacking learning and semi-supervised classification (SSC-SL), and to achieve high prediction accuracy with this method in the case of few soil sampling points.

2. Materials and Methods

2.1. Study Area

The study area is located in the northern part of Jurong City, Jiangsu Province, China (115°45′22″ to 117°14′44″E and 31°16′16″ to 29°47′3″), on the south side of the lower reaches of the Yangtze River, with an area of 50.46 km^2 (Figure 1). The study area is located in the subtropical monsoon climate zone, with a mild climate, favourable temperatures, moderate rainfall, and sufficient light. The geomorphology of the study area is mainly hilly, with an overall distribution of north–south-trending bar-shaped hills, large changes in elevation, and distinct topographic relief [21]. In the northeastern part of the study area, there is the Lunshan Reservoir and Gaoli Mountain [22], and the total elevation of the study area ranges from 28 m to 403 m.

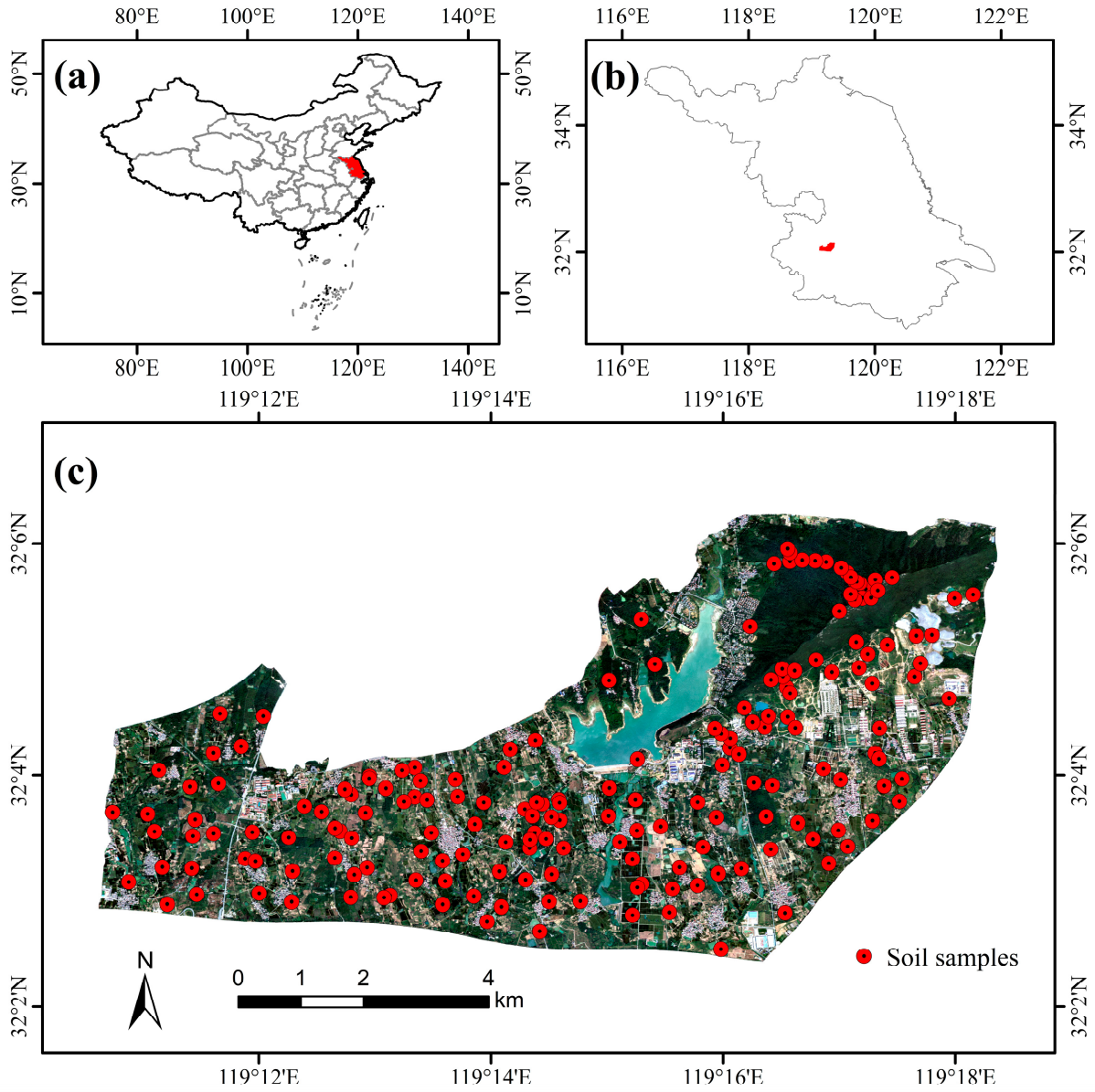

Figure 1. (**a**,**b**) Location of the study area; (**c**) distribution of sampling points and remote sensing image.

The land used in the area is primarily forest lands, followed by cultivated lands and gardens. The forest lands are widely distributed, with their natural vegetation mainly consisting of subtropical evergreen broadleaf forests and shrublands, typically found on the mid and upper slopes. The forest lands are most extensively spread in Gaoli Mountain. In the paddy fields, the artificially cultivated crops are mainly rice and wheat, distributed in a north–south-oriented strip pattern along the valleys. Drylands and orchards are generally located in the middle parts of the hills. The soil parent materials include light purple sandstone, Quaternary red clay, and other soil parent materials. Light purple sandstone is sparsely distributed in the central part of the study area, while Quaternary red clay is found in the middle of the southern slopes of Gaoli Mountain, with the remainder being other soil parent materials. The type of land use, topography, and soil parent material influence or reflect the formation and distribution of soils in the study area, with the specific spatial distribution detailed in Figure 2.

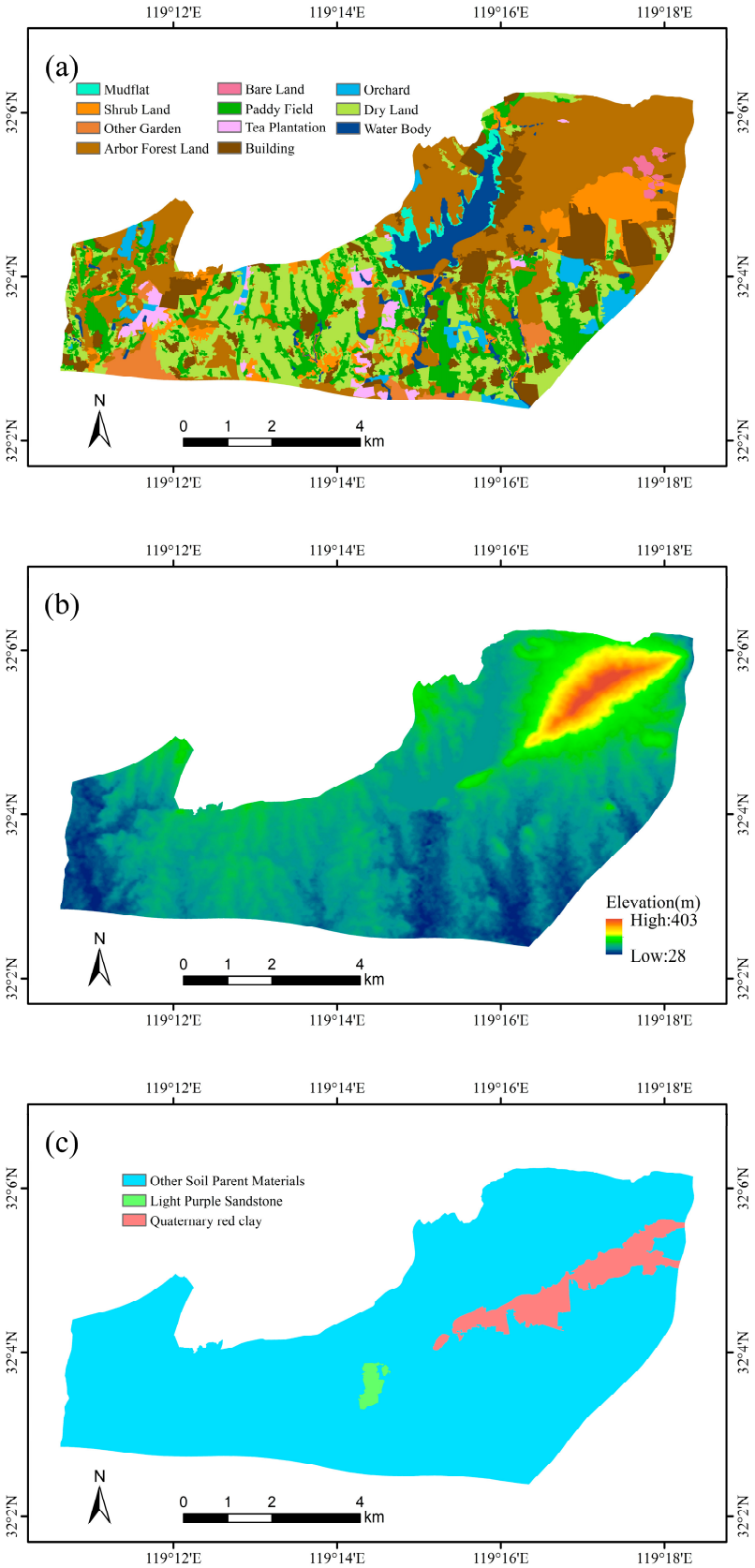

Figure 2. Main environmental variables in the study area: (**a**) map of land use; (**b**) map of DEM; (**c**) map of soil parent material.

2.2. Soil Data

The soil sample collection followed the principles of uniformity and representativeness to ensure that all sampling points were relatively uniform in spatial distribution, while single sampling points represented the same or similar landscapes within a certain range, i.e., having the same or similar topography, soil parent material, and land use. In addition, all sampling points covered different landscapes within the study area. We collected 183 soil profile samples (0–120 cm) from March 2021 to November 2022. We identified 10 soil subgroups according to the Chinese Soil Taxonomy [23], which are Typic Purpli-Udic Cambosols, Typic Hapli-Udic Cambosols, Typic Hapli-Udic Argosols, Typic Fe-accumuli-Stagnic Anthrosols, Typic Claypani-Udic Argosols, Red Ferri-Udic Cambosols, Red Ferri-Udic Argosols, Mottlic Hapli-Udic Argosols, Lithic Udi-Orthic Primosols, and Endogleyic Fe-accumuli-Stagnic Anthrosols. The spatial distribution of the soil sampling points is shown in Figure 1.

2.3. Environmental Variables Data

Soil type is influenced by five main factors: climate, topography, biology, time, and parent material [24]. However, in smaller study areas, common environmental variables used for larger regions, such as macroclimate or annual average rainfall, show limited variation and, thus, are not applicable. Given this, and considering the relationship between the soil types and environmental variables in our study area, we obtained data on two types of environmental variables (topographic data and remote sensing data) from existing sources and generated some derived environmental variables. In addition, the soil parent material data were obtained from actual investigations based on historically retained geologic maps. The definitions of the three types of environmental variables (including derived environmental variables) are given in Table 1.

Table 1. Selected set of environmental variables.

Type	Name of Environmental Variables/Definition	Abbreviation
Remote Sensing	**Simple Ratio (NIR/R)**	**SR**
	Normalised Difference Vegetation Index ((NIR − R)/(NIR + R))	**NDVI**
	Green Ratio Vegetation Index (NIR/G)	**GRVI**
	Normalised Difference Water Index ((G − NIR)/(G + NIR))	**NDWI**
	Green Leaf Index ((2 ×G − B −R)/(2 × G + B + R))	**GLI**
	Land Use	**LU**
Terrain	**Elevation (m)**	**Ele**
	Aspect	Asp
	Slope	**Slo**
	Plan Curvature	PlC
	Profile Curvature	PrC
	Topographic Position Index	TPI
	Topographic Wetness Index	**TWI**
	Multi-Resolution of Ridge Top Flatness Index	**MRRTF**
	Multi-Resolution Valley Bottom Flatness Index	**MRVBF**
	Horizontal Distance to Ridge Line	**HDRL**
	Horizontal Distance to Valley Line	**HDVL**
Soil Parent Material	**Soil Parent Material**	**SPM**

NIR: near-infrared band of GF-2; R: red band of GF-2; G: green band of GF-2; B: blue band of GF-2.

2.3.1. Remote Sensing Data

In related studies, various bands and derivatives of remote sensing images have been proposed to serve as environmental variables for soil type mapping [25] and to be effective in improving the prediction of soil classification [26]. We downloaded the Gaofen-2 (GF-2) image data from the China Centre for Resources Satellite Data and Application (https://data.cresda.cn/#/home, accessed on 2 December 2023). We chose the image period of low

cloudiness during the crop flowering–irrigation period; the feature information is clearer, and the resulting feature reflectance spectral differences are more significant and easier to distinguish and identify. The GF-2 image includes panchromatic images (0.45–0.90 μm) and multispectral images (near-infrared band: 0.77–0.89 μm; red band: 0.63–0.69 μm; green band: 0.52–0.59 μm; blue band: 0.45–0.52 μm). The spatial resolution of the multispectral images is 1 m, while that of the panchromatic images is 4 m. The images were captured on September 19, 2019. The following pre-processing steps were performed on the GF-2 image: (1) multi-spectral image: radiometric calibration, atmospheric correction, ortho-rectification; (2) panchromatic image: radiometric calibration, ortho-correction; (3) image fusion of multi-spectral image and panchromatic image. Based on the fused images, we calculated the following vegetation indices as environmental variables, including Simple Ratio (SR), Normalised Difference Vegetation Index (NDVI), Green Ratio Vegetation Index (GRVI), Normalised Difference Water Index (NDWI), and Green Leaf Index (GLI).

Additionally, based on remote sensing image data, we initially utilised the random forest model to obtain a pre-processed land use map (LU). To prevent the excessive fragmentation of patches, we merged those smaller than 600 m^2 and manually adjusted their boundaries. The revised land use maps were subsequently incorporated into our model as environmental variables. Given the impact of land use classification features on soil classification accuracy, we refined the land use types within the study area as much as possible. We divided the land use in the study area into 11 categories, including cultivated land (paddy field and dry land), forest land (arbour forest land and shrub land), garden (tea plantation, orchard, and other gardens), mudflat, bare land, water body, and building. For instance, paddy fields were considered for distinguishing Anthrosols from other soil types. Arbour forest land and shrub lands can be used to differentiate between Argosols and Cambosols, while gardens help to distinguish the subcategories within Argosols. Tidal flats and bare lands usually reflect soil types that are categorised into the main soil types near the patches.

2.3.2. Terrain Data

Terrain factors are a major important factor affecting the variation in soil types. The change in topography affects the change in the spatial distribution of soil types [27]. A more accurate map of soil types can be obtained after adding topographic attributes [28]. We downloaded the ALOS digital elevation model (DEM) data from the Alaska Satellite Facility (https://search.asf.alaska.edu/, accessed on 2 December 2023) with a resolution of 12.5 m and calculated the following topographic metrics in SAGA-GIS (version 8.4.1): aspect, slope, plan curvature (PlC), profile curvature (PrC), Topographic Position Index (TPI), and Topographic Wetness Index (TWI). In addition, we additionally calculated two other topographic indicators: horizontal distance to ridge line (HDRL) and horizontal distance to valley line (HDVL).

2.3.3. Soil Parent Material Data

Soil parent material is the basis of soil formation and plays an important role in soil classification [29]. Soil parent material data have been used as one of the environmental variables to predict soil types in many studies investigating soil classification [28,30]. We roughly determined the distribution boundaries of the soil parent material based on historical geology maps. Finally, the soil parent material map was obtained from field surveys.

2.3.4. Pre-Processing of Environmental Variables

We conducted one-hot encoding for incorporating the factor-type data (SPM and LU) into the modelling analysis. We filtered the appropriate environmental variables from the remaining environmental variables for modelling analysis. The importance of the remaining environmental variables was calculated without using any specific model, and this served as the basis for selecting suitable environmental variables. This analysis

was implemented using the filterVarImp function of the "caret" package in R [31]. We analysed each environmental variable using the ROC curve, using the area under the curve as the score, and deleted the environmental variables with a weighted average score of 0.85 or less; the deleted environmental variables included Asp, PlC, PrC, and TPI. The remaining 12 environmental variables were obtained as SR, NDVI, GRVI, NDWI, GLI, Ele, Slo, TWI, MRRTF, MRVBF, HDRL, and HDVL. Finally, the above 14 environmental variables were selected as the environmental variables involved in the modelling (environmental variables in bold in Table 1). To ensure consistency in the spatial resolution of the various environmental variables, as well as to maintain good analytical and operational efficiency, we employed the nearest neighbour method to resample the environmental variables to a 2 m spatial resolution. Figure 2 shows the main environmental variables in the study area.

2.4. Stacking Learning (SL)

Machine learning methods are widely used in digital soil mapping, but the dominant approaches today use a single machine learning method to predict soil properties or types. Ensemble learning is an important part of the machine learning field, which includes three different methods, namely, bagging, boosting, and stacking learning techniques [32]. Stacking learning is a branch of ensemble learning. It was first proposed by Wolpert in 1992 [33]. It combines different base learners by constructing a meta-model, which can be any type of model, e.g., decision tree, neural network, or linear regression. Typically, stacking learning can achieve high accuracy because it has the advantage of combining multiple base learners in a weighted way, which can compensate for the shortcomings of a single base learner in model prediction and improve the overall prediction performance. Figure 3 shows that stacking learning can achieve more accurate prediction results.

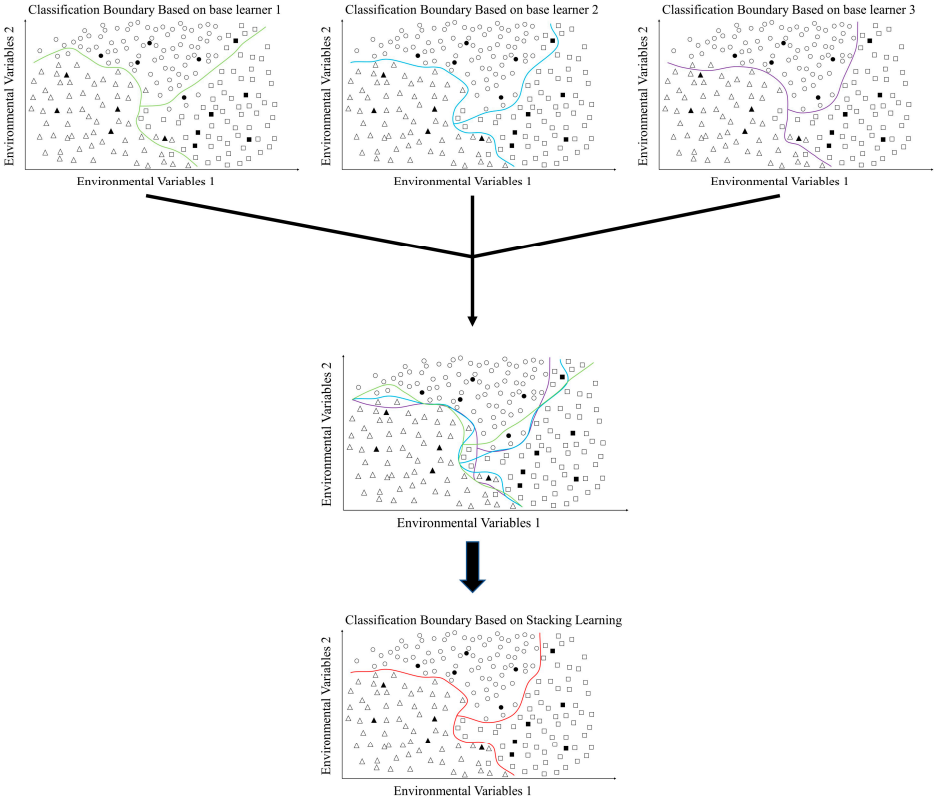

Figure 3. Stacking learning schematic. Black triangles, squares and circles indicate environmental variables at the actual sampling point location. White triangles, squares and circles indicate environmental variables at other locations. Lines of different colors indicate classification boundaries generated according to different models.

2.4.1. Ranger

The Ranger model is a fast implementation of the random forest model, which is more computationally efficient than traditional random forests. Ranger trains and analyses the data using a large number of decision trees, and a subset of features is randomly selected from the full set of features when training each tree [34]. Each tree is trained on a slightly different set of data, and the final prediction is made by voting or averaging the predictions from all of the decision trees. In this way, the model can handle a large number of features and has good generalisation capabilities.

2.4.2. Rpart

The Rpart model generates a decision tree by recursively partitioning the data. At each segmentation stage, the model searches for the split variables and the best split points to minimise the uncertainty or variance of the target variables in the nodes. This process continues until certain stopping criteria are met, such as the maximum depth of the tree or a minimum number of samples per node [35]. A notable advantage of this model is its ease of understanding and explanation. It allows for clear visibility into which features are pivotal in the prediction process and how the values of these features impact the predicted outcomes.

2.4.3. XGBoost

XGBoost uses a gradient boosting framework, developed by Tianqi Chen [36] in 2014. In this framework, the model is improved incrementally by adding new decision trees, with each new tree attempting to correct the prediction errors of the previous tree. This approach helps the model to fit the data better and usually results in better prediction performance than the decision trees alone. XGBoost makes many improvements on the basis of gradient boosting, including regularisation, the parallelisation of computation, and the automatic handling of missing values, in order to improve the accuracy and computational efficiency of the model. In addition, the efficiency and flexibility of XGBoost make it a good choice for large-scale data and complex prediction problems.

2.4.4. Model Parameter Setting

We used the stacking learning method to combine three base learners to construct an ensemble learning model for predicting soil subgroups in the study area. The three base learners are Ranger, Rpart, and XGBoost. We set the hyperparameters for each base learner. The Ranger model adjusted the two parameters of mtry and num.trees; the Rpart model used the default parameters; and the XGBoost model adjusted the five parameters of booster, eta, max_depth, subsample, and colsample_bytree. The specific parameters given to each base learner are shown in Table 2. The models were implemented in R (version 4.0.2) [37] and RStudio (version 1.3.1093) [38] using the sl3 package [39]. The sl3 package implements the stacking learning algorithm proposed by van der Laan et al. [40]. It integrates a set of base learners and assigns the weights of each base learner by the minimum risk value from cross-validation to obtain the optimal model. According to the results of Van der Laan [40] and others, the algorithm is robust and the accuracy of the results obtained is usually better than that of the best single base learner.

Table 2. Specific parameters of the three base learners.

Base Learner	Definition	Hyperparameters	Hyperparameter Values	References
Ranger	A fast implementation of random forests	mtry num.trees	4 500	[34]
Rpart	Recursive partitioning and regression trees	defaults	-	[35]
XGBoost	Extreme gradient boosting	booster eta max_depth subsample colsample_bytree	gbtree 0.3 6 1 1	[36]

2.5. Selection of Unlabelled Sample Points

We used the fuzzy c-means algorithm (FCM) for the cluster analysis of the environmental variable factors. FCM was proposed by Dunn in 1973 [41]. Compared to traditional clustering algorithms, the advantage of FCM is that it is not directly and rigidly assigned to a particular clustering category; rather, the relationship between each data point and each clustering category is expressed as a weight, which means it can be assigned to more than one clustering result, giving the clustering results a high degree of robustness, especially when there is noise in the data.

The FCM algorithm can control the extent of multiple clustering categories corresponding to each data point by setting the degree of fuzziness. Among the many clustering algorithms, the FCM algorithm has been widely and successfully applied. For example, Li et al. [42] used FCM to classify annual mean soil temperatures at 20 cm from 141 ST observation sites on the Tibetan Plateau from 1981 to 2020, and Peng et al. [43] used FCM to map nine Danish terron classes. As soil types and landscapes are closely related, changes in the landscape may reflect changes in soil types. When mapping soil types, the same soil type often has the same landscape. The purpose of using FCM prior to applying the semi-supervised classification method is to delineate a clustering result to determine the location of the added unlabelled sample points in order to obtain more accurate map plot boundaries for soil type mapping. In a similar study, Dunkl et al. [44] used environmental variables to cluster the landscapes within the study area into spatial units and used them to help delineate soil texture in predictive mapping.

We constructed the FCM using the geocmeans package [45] in R and determined the clustering category to which each soil sample point belonged. We set the number of clusters (k) from 10 to 20 with a step of 1, and the degree of fuzziness (m) from 1.1 to 2 with a step of 0.1. The maximum value of Silhouette was used to determine the corresponding values of k and m (Figure 4), which ultimately determined k to be 17 and m to be 2. After merging the areas where the patch area is less than 600 m^2, we finally obtained 14 clustering categories (Figure 5).

$$s(i) = \frac{b(i) - a(i)}{(a(i), b(i))}, \quad (1)$$

where $s(i)$ ranges from -1 to 1, $a(i)$ denotes the similarity between object i and other objects in the cluster, and $b(i)$ denotes the similarity between object i and all objects in the nearest cluster.

2.6. Semi-Supervised Classification (SSC)

Traditional digital soil mapping methods typically require a large amount of labelled soil sample data, and the purpose of soil mapping is achieved by fitting the non-linear relationship between environmental variables and soil samples [46]. However, collecting a large number of soil samples requires significant human and material resources, as well as considerable time. Semi-supervised classification, an important branch of machine learning, is particularly effective for datasets with a small number of samples. This method relies

on a limited amount of labelled data combined with self-training on unlabelled data. This approach helps to address the challenge of small sample sizes and ultimately enhances the predictive performance of the model [47]. Figure 6 shows that more accurate classification boundaries can be obtained after adding unlabelled samples using clustering methods.

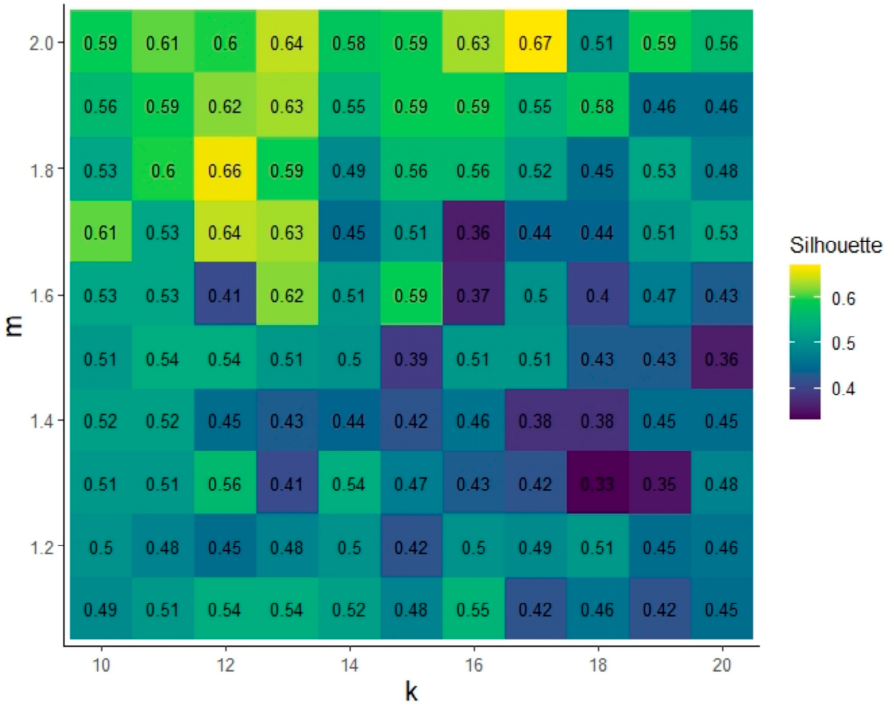

Figure 4. Silhouette values for different m and k.

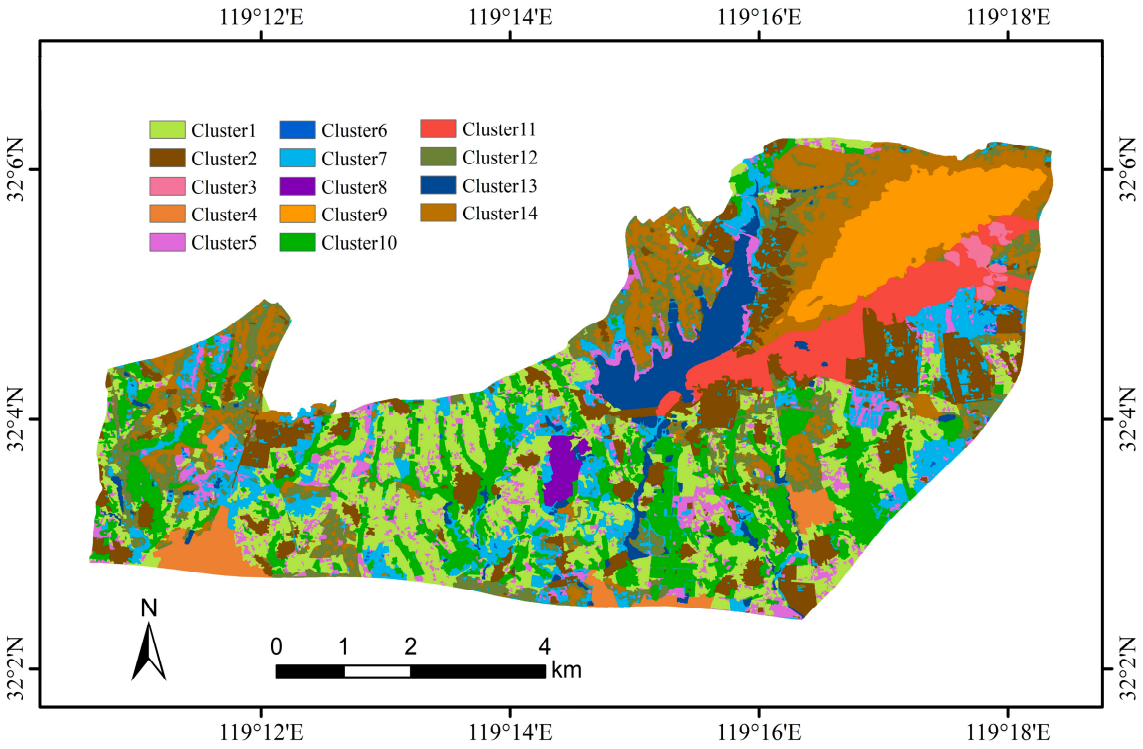

Figure 5. Map of FCM.

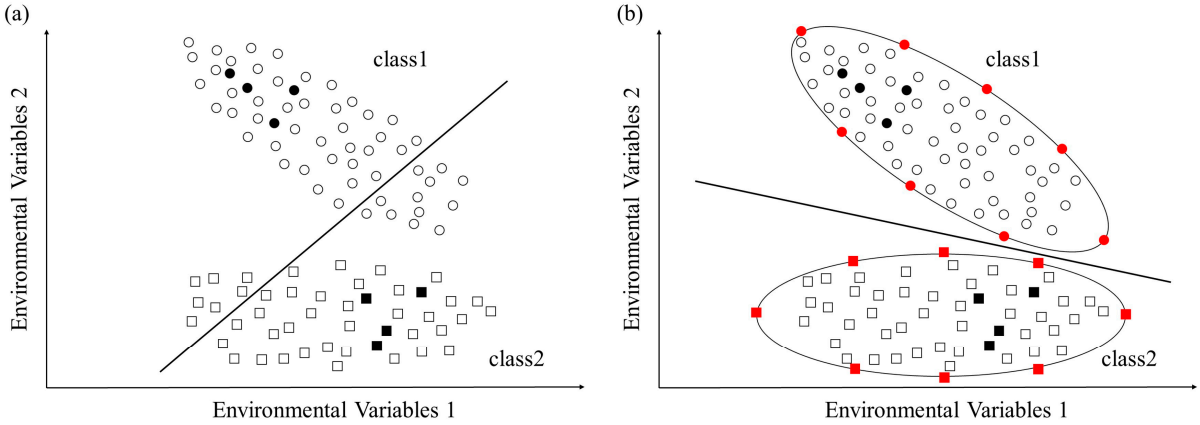

Figure 6. Schematic of semi-supervised classification: (**a**) boundary obtained using only the original soil samples; (**b**) more accurate boundary obtained by applying semi-supervised classification with the addition of unlabelled samples using FCM. Black squares and circles indicate environmental variables for original soil sample locations. Red squares and circles indicate environmental variables for unlabeled samples. White squares and circles indicate environmental variables at other locations.

We selected one unlabelled sample point every 100 m at 15 m intervals from the boundary of each patch in the cluster map obtained using FCM and obtained a total of 8703 unlabelled sample points. We used the random forest as a self-training learner for semi-supervised classification by first training the labelled sample point data and then predicting all unlabelled sample points. Then, the unlabelled sample points with a confidence threshold of 0.9 were added to the training set and retrained to generate the four models: SSC-SL, SSC-Ranger, SSC-Rpart, and SSC-XGBoost. The SSLR package [48] was used to implement the semi-supervised classification models in R.

2.7. Model Prediction

The research method of this paper is as follows (Figure 7):

(1) Ranger, Rpart, and XGBoost were selected as the base learners to model and predict the unlabelled data. The optimal tuning of the model hyperparameters and modelling prediction on unlabelled data were carried out. Finally, the prediction results of each of the three base learners were obtained.
(2) Using random forest as a meta-learner, the prediction results of the three base learners were combined to obtain the stacking learning model.
(3) The cluster analysis of the environmental variables was performed using the FCM model to generate a cluster map, and then the unlabelled dataset was selected near the patch boundaries of the cluster map.
(4) An initial model was trained with labelled data and then the unlabelled dataset was predicted.
(5) We established a confidence threshold range from 0.55 to 0.95, with increments of 0.05. Within this range, we filtered the unlabelled datasets that met or exceeded a specific confidence threshold and subsequently merged them with the labelled data.
(6) Steps 4–5 were repeated until no new data reached the preset confidence threshold, and the final training set, extended by the semi-supervised classification method, was obtained.
(7) The training set expanded by the semi-supervised classification method in the SL model and its three base learners were remodelled to obtain the four models of SSC-SL, SSC-Ranger, SSC-Rpart, and SSC-XGBoost, and the original four models (SL, Ranger, Rpart, and XGBoost) were outputted for comparison.
(8) The soil map was predicted using eight models.

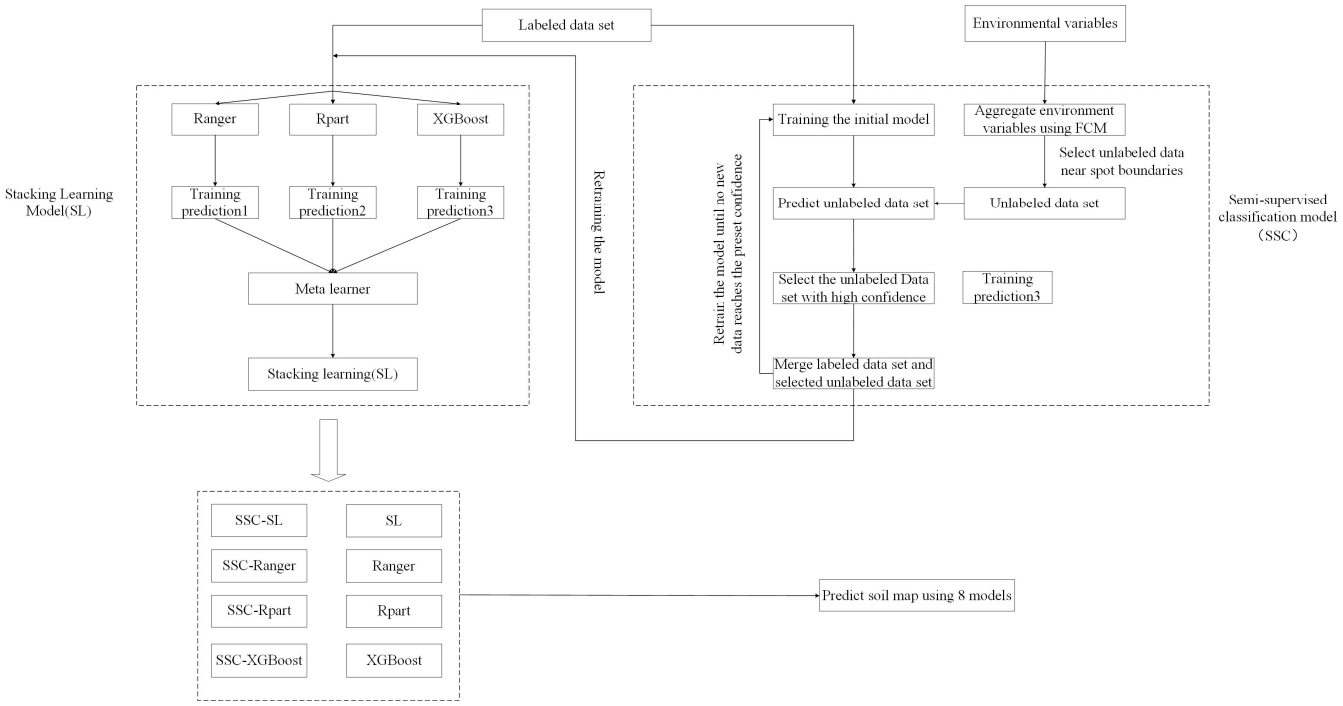

Figure 7. Overall framework of the research method.

2.8. Assessment of Model Accuracy

To assess the accuracy of the soil mapping, we randomly divided the soil sampling points into two parts, selecting 70% of the soil sampling points as training sampling points and 30% of the soil sampling points as validation sampling points. In this study, two evaluation metrics were calculated for 100 repeats of 5-fold cross-validation to analyse the model performance: overall accuracy (OA) and kappa coefficient. Higher values of OA and kappa coefficient indicate better classification.

2.9. Importance Analysis of Environmental Variables

In order to explore the importance of each environmental variable in each model, by removing an environmental variable and keeping the remaining environmental variables unchanged, the difference in the negative log-likelihood obtained by removing the variable and not removing the variable was compared, with a larger difference indicating that the variable was more important in the model's predictions.

3. Results

3.1. Prediction Accuracy of Different Models

Through the validation of 52 sampling points, the accuracy of eight different models was assessed, yielding overall accuracy (OA) and kappa coefficients for each (as shown in Figures 8 and 9). Figures 8 and 9 demonstrate that the SSC-SL model, integrating semi-supervised and stacking learning methods, achieved the best predictive performance among the eight models. The SSC-SL model's optimal OA and kappa were 0.77 and 0.73, respectively, surpassing the three base models (Ranger, Rpart and XGBoost) by margins of 2.8%, 21.1%, and 2.8% for OA, and 3.0%, 25.8%, and 3.0% for kappa. Within the SL model's base learners, the performance ranking from highest to lowest was Ranger, XGBoost, and Rpart. The SL model, by combining the strengths of these three base learners, achieved superior predictive performance with OA and kappa coefficients of 0.71 and 0.66, respectively. The OA of the SL model exceeded the three base learners by 2.8%, 21.1%, and 2.8%, while its kappa coefficients were 4.7%, 4.7%, and 36.7% higher.

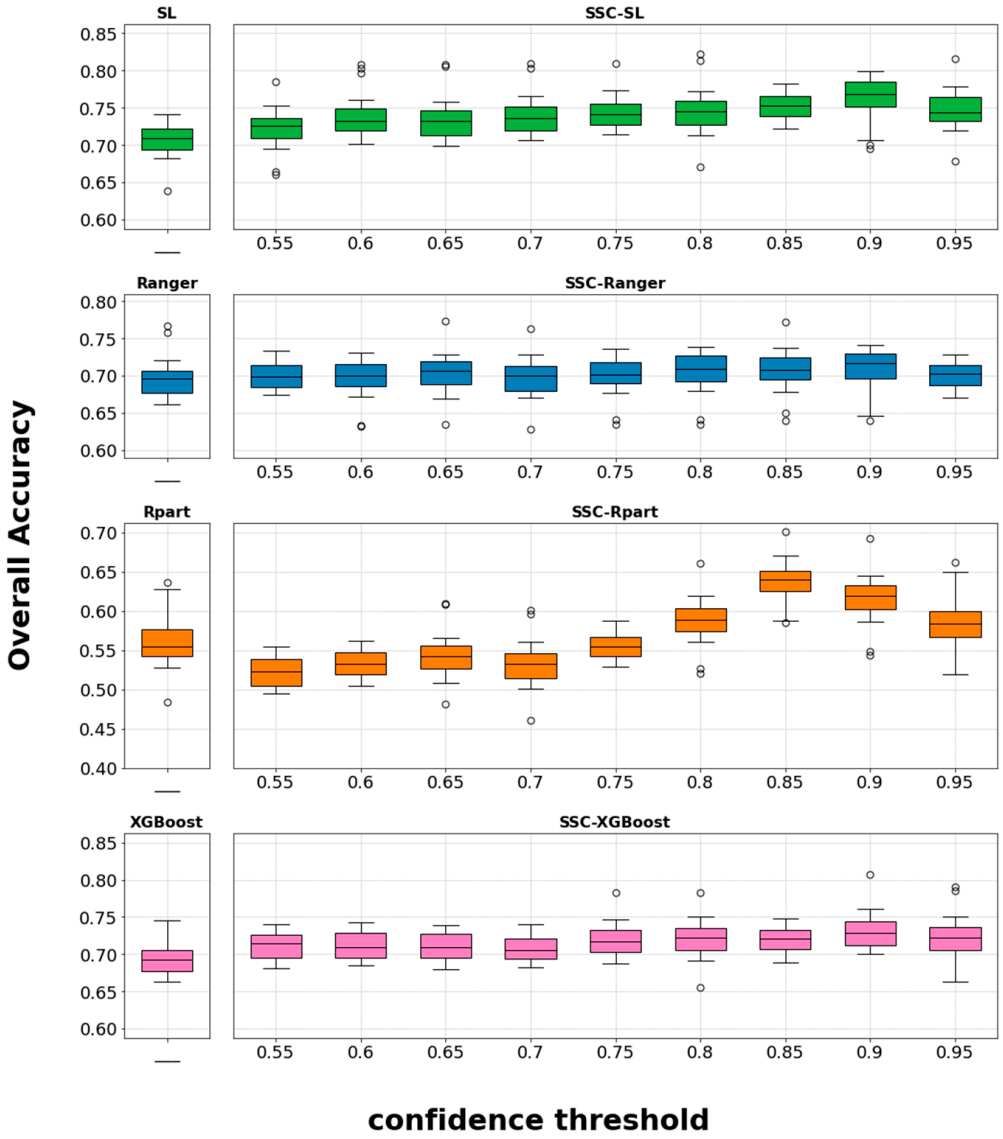

Figure 8. OA values of different models.

When integrated with semi-supervised classification methods, the models SSC-SL, SSC-Ranger, and SSC-XGBoost consistently outperformed their counterparts SL, Ranger, and XGBoost across various threshold settings in terms of predictive effectiveness. However, at thresholds between 0.55 and 0.7, the performance of SSC-Rpart was lower than that of the Rpart model. The optimal confidence level for SSC-SL, SSC-Ranger, and SSC-XGBoost was 0.9, while for SSC-Rpart, it was 0.85. At these optimal thresholds, the predictive performances of SL and the three base learners improved to some extent.

When SSC-SL, SSC-Ranger, SSC-Rpart, and SSC-XGBoost were set at their optimal confidence thresholds, their overall accuracy (OA) rates improved by 7.8%, 2.8%, 9.7%, and 5.5%, respectively, compared to scenarios where semi-supervised classification methods were not integrated. The kappa coefficients also showed enhancements of 9.6%, 3.0%, 12.5%, and 7.2%, respectively. In terms of the enhancement of predictive accuracy, the SSC method's effectiveness for various foundational learners was ranked from highest to lowest as Rpart, SL, XGBoost, and Ranger. The SSC-SL model, in particular, offered the most consistent high-precision predictions, proving to be a viable approach for improving the accuracy of digital soil mapping.

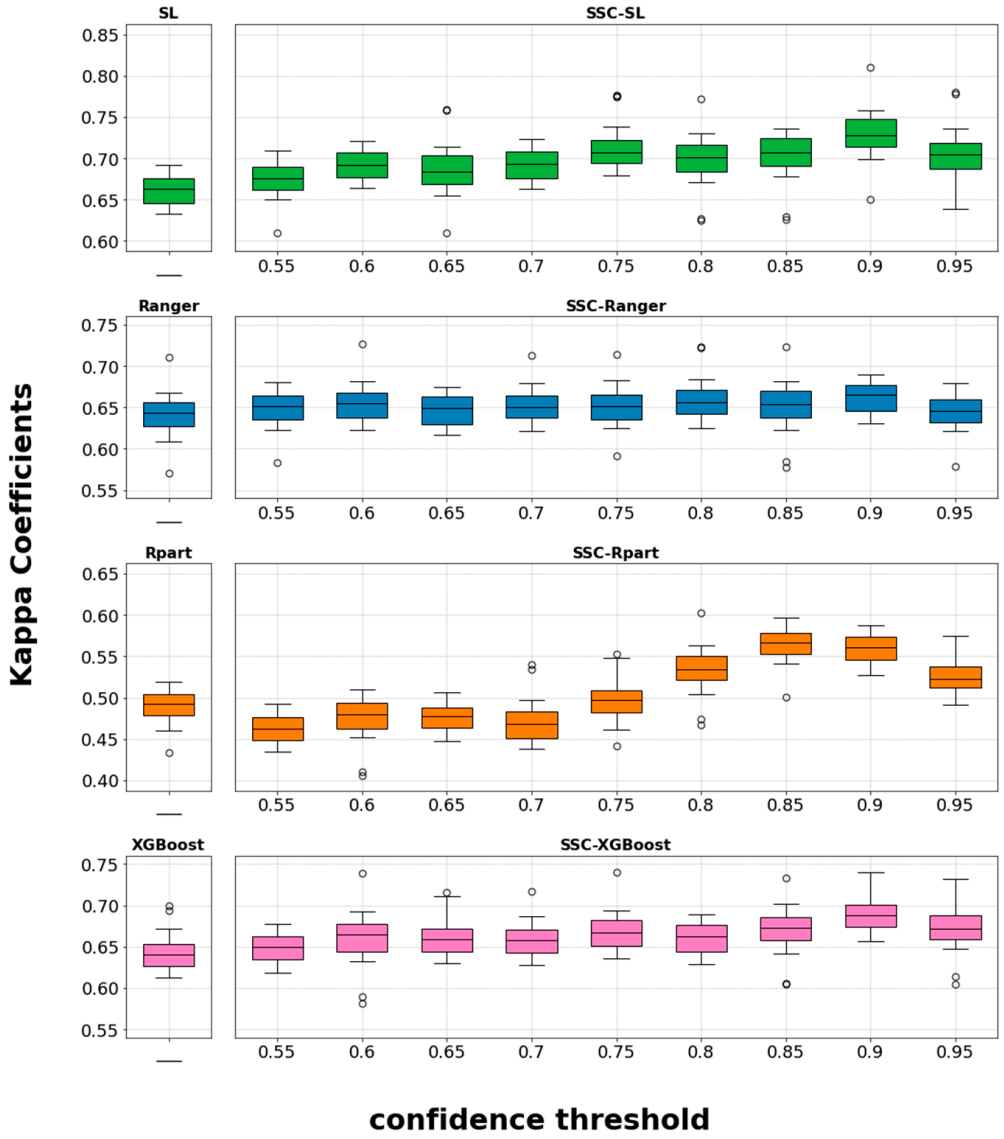

Figure 9. Kappa coefficients of different models.

3.2. Spatial Distribution of Soil Subgroups

Figure 10 presents the soil maps generated by SL, Ranger, Raprt, and XGBoost, and their semi-supervised counterparts SSC-SL, SSC-Ranger, SSC-Raprt, and SSC-XGBoost at their respective optimal confidence threshold. We can see that the soil maps generated by the different models are generally consistent. Lithic Udi-Orthic Primosols are found at the summit of Gaoli Mountain in the northeastern part of the study area, where the higher elevation and weaker soil development led to the formation of this soil subgroup. As the elevation decreases, the soil develops gradually and Typic Hapli-Udic Cambosols are formed on the upper and middle slopes of Gaoli Mountain, which are also distributed in small quantities in the upper regions of other hills in the study area. Red Ferri-Udic Cambosols and Red Ferri-Udic Argosols are also formed in the middle and lower parts of the southern slope of Gaoli Mountain, which are two subgroups of soils formed by the weak and strong development of the Quaternary red clay. In the inter-slope valleys of the study area, Endogleyic Fe-accumuli-Stagnic Anthrosols and Typic Fe-accumuli-Stagnic Anthrosols have been formed due to the long-term artificial cultivation of rice. The bottom of the Endogleyic Fe-accumuli-Stagnic Anthrosols is affected by groundwater and remains in a reduction state all year round, and its distribution area is more low-lying than Typic Fe-accumuli-Stagnic Anthrosols. A small amount of Typic Purpli-Udic Cambosols is found

in the central part of the study area, and this small area of soil is due to soil parent material. The formation of Mottlic Hapli-Udic Argosols requires some accumulation of soil clay particles and redox reactions within the soil layer. This means that a combination of groundwater and surface water infiltration is required, which is why this soil subgroup is mainly found in arid areas on the lower and middle parts of slopes. Typic Hapli-Udic Argosols do not require redox reactions within the soil layer and are therefore mainly found in dryland areas in the upper and middle parts of slopes. Typic Claypani-Udic Argosols have a smaller area and are mainly interspersed between Mottlic Hapli-Udic Argosols and Typic Hapli-Udic Argosols.

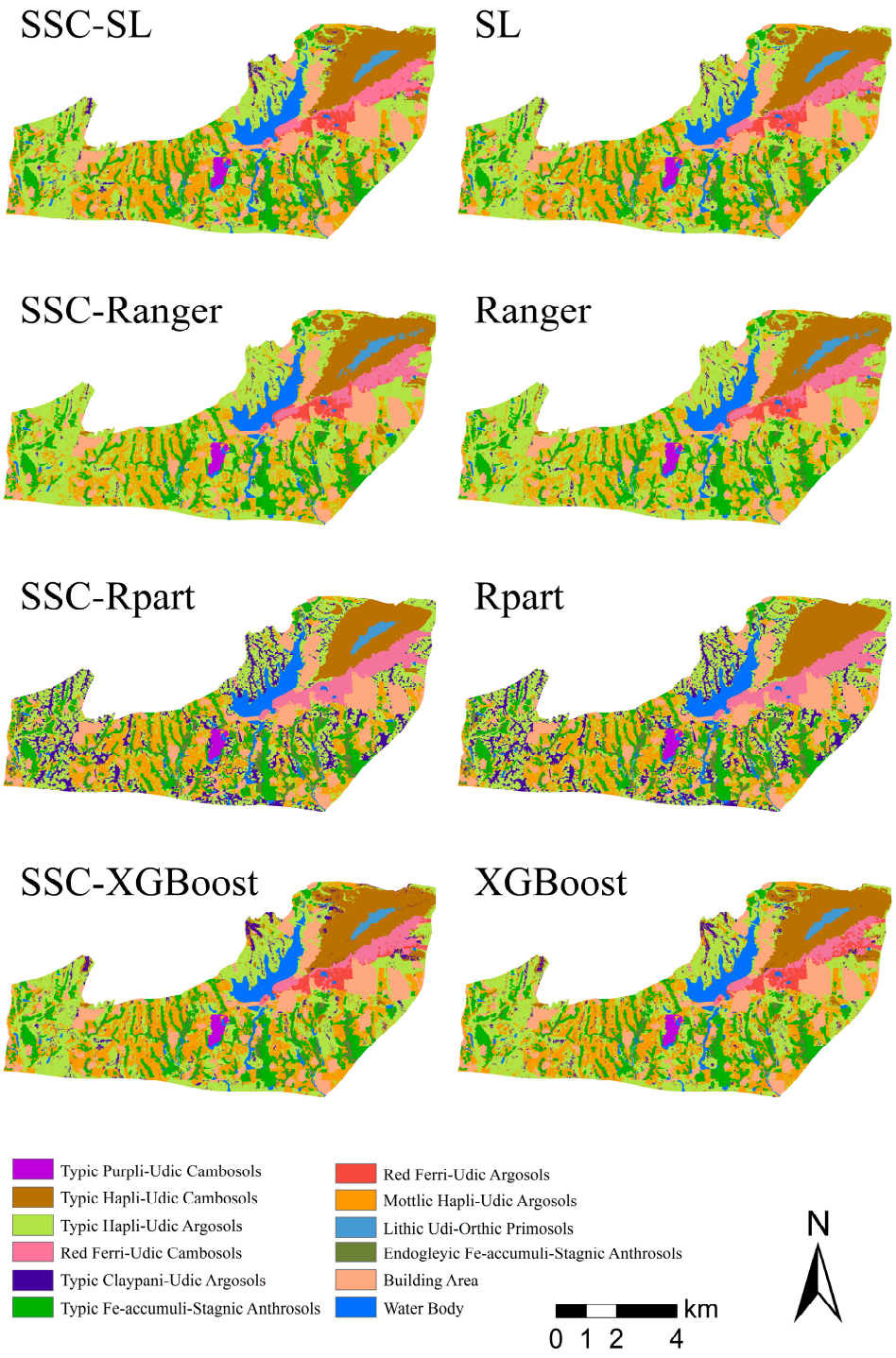

Figure 10. Spatial distribution map of soil subgroups from different models.

Comparing the distribution of soil subgroups obtained by three learners and the SL method, we can observe that the distribution of soil types obtained by the SL method is more reasonable compared to the base learners. Additionally, the soil subgroup distributions obtained using the Ranger and XGBoost methods are notably superior to those obtained using Rpart. For instance, in the Rpart model, the distribution of Typic Claypani-Udic Argosols is excessively broad, whereas in Ranger and XGBoost, it is more constrained, and SL presents a more accurate range. Similarly, while the Rpart model fails to predict Red Ferri-Udic Argosols correctly, SL, integrating the strengths of Ranger, XGBoost, and Rpart, achieves accurate predictions of this soil type. Additionally, soil spatial distributions derived from the same category of machine learning methods exhibit similar characteristics, as evidenced by the notable similarity between Ranger and SSC-Ranger compared to Rpart and XGBoost. Moreover, machine learning models employing semi-supervised classification methods reveal more detailed content at the boundaries of soil category distributions, aligning more closely with the actual subgroup distributions. For instance, SSC-SL builds upon the foundation of SL by providing a clearer delineation of the details of Endogleyic Fe-accumuli-Stagnic Anthrosols. It more accurately depicts the spatial distribution characteristics of this soil type compared to SL.

3.3. Importance Analysis of Environmental Variables

Figure 11 shows the importance ranking of each environmental variable for SL, Ranger, Raprt, and XGBoost and their semi-supervised counterparts, SSC-SL, SSC-Ranger, SSC-Raprt, and SSC-XGBoost at their respective optimal confidence threshold. As can be seen from Figure 11, the importance ranking of each model did not change much after the addition of the semi-supervised classification method. There were only a few changes in the ranking positions of the importance ranking of each environmental variable.

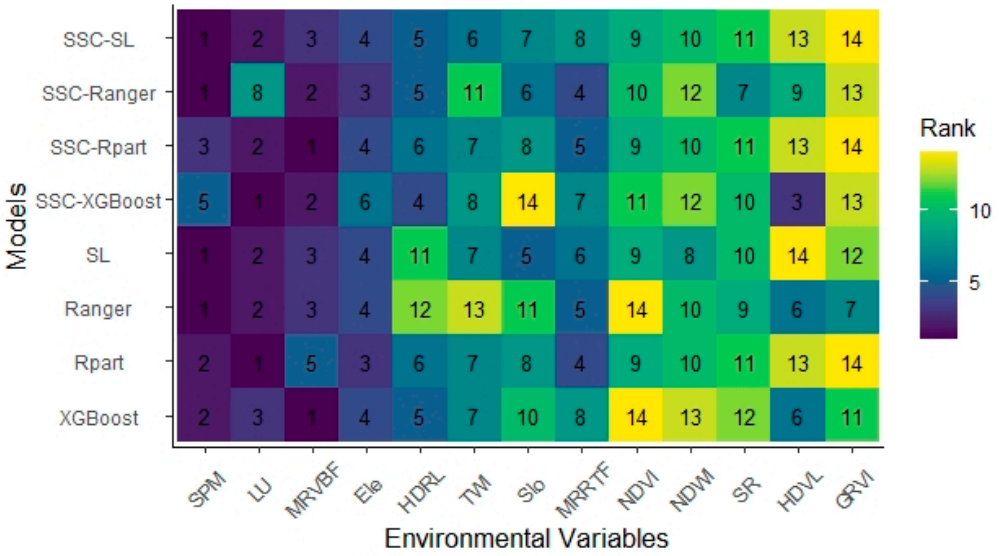

Figure 11. Ranking of environmental variables' importance in different models.

SPM, LU, MRVBF, and Ele have high importance rankings in each of the eight models. SPM is the basis of soil formation and is the first influence on soil formation. For instance, it can be employed here to distinguish between Red Ferri-Udic Cambosols, Red Ferri-Udic Argosols, Typic Purpli-Udic Cambosols, and other types of soils. LU identifies soil subgroups by the difference in the type of land use in which they are located. For example, Endogleyic Fe-accumuli-Stagnic Anthrosols and Typic Fe-accumuli-Stagnic Anthrosols are generally found in paddy fields, while Typic Hapli-Udic Argosols are found in dry lands or forested areas. MRVBF and Ele reflect the changes in the soil subgroups from the perspective of terrain variation. For instance, Typic Hapli-Udic Cambosols and Lithic Udi-Orthic Primosols exhibit distinct characteristics: the former shows lower MRVBF

values and higher Ele values, while the latter demonstrates higher MRVBF values and lower Ele values.

4. Discussion

In this study, we innovatively apply the combination of semi-supervised classification and stacking learning methods in a model named SSC-SL for predicting soil types. Our results demonstrate that the SSC-SL model outperforms seven other methods in terms of prediction accuracy. In particular, the combination of SL and SSC-SL assembles their respective base learners and leverages their strengths to achieve a higher prediction accuracy. This finding aligns with the research outcomes of Taghizadeh-Mehrjardi and Tao et al., who asserted that SL can achieve a higher level of accuracy than any single base learner [19,49]. Other researchers, such as Amin Sharififar et al. [50] created a soil map for the North Bikomi District in North Central Timor, East Nusa Tenggara Province, using only KNN, RF, and SVM to compare their OA, which ranged between 0.55 and 0.75. Tarek Assami et al. [51] produced a soil map for the Zeb El Gherbi region in the southern piedmont of the Saharan Atlas Mountains in Southeast Algeria, employing six machine learning models: bagged classification tree, random forest, linear support vector machines, radial basis support vector machines, single-hidden-layer neural networks, and multilayer perceptron neural network. They compared their kappa coefficients, which varied from 0.38 to 0.47. The accuracy of their cartographic results was largely dependent on the chosen machine learning models.

Furthermore, the SSC-Ranger and SSC-XGBoost models demonstrate overall consistent predictive performance, characterised by an initial increase followed by a decrease as the confidence level rises, peaking at optimal predictive efficiency when the confidence level reaches 0.9. The predictive capabilities of both the SSC-Ranger and SSC-XGBoost models are enhanced across various confidence levels via semi-supervised classification methods, albeit to varying degrees. The SSC-Ranger model exhibits a smaller extent of improvement in predictive performance. Numerous studies have shown that Ranger is a relatively accurate model, making significant precision improvements in SSC-Ranger challenging. This aligns with the findings by Zhang et al. [20] regarding the performance of random forests in semi-supervised classification models. On the other hand, the predictive performance of SSC-XGBoost is comparable to SSC-Ranger, possibly due to SSC-XGBoost's higher accuracy in the unlabelled dataset compared to Ranger, which contributes to its more substantial precision enhancement. Meanwhile, the SSC-Rpart model outperforms the Rpart model in predictive accuracy at confidence levels ranging from 0.7 to 0.9, while it underperforms at confidence levels between 0.55 and 0.7. This could be attributed to the relative simplicity of the Rpart model and its lower predictive precision. At lower confidence levels, the generated unlabelled dataset has lower accuracy, leading to the inclusion of a significant amount of incorrect pseudo-label data, resulting in a decline in predictive performance, even falling below that of the Rpart model itself. Additionally, due to the lower accuracy of the Rpart model, SSC-Rpart is likely to achieve a greater degree of precision improvement at the optimal threshold. SSC-SL maintains a stable enhancement in predictive performance across various confidence levels. SSC-SL is a robust model that combines the features of SSC-Ranger, SSC-Rpart, and SSC-XGBoost, effectively mitigating the risk of reduced predictive performance experienced when using a single model, such as the SSC-Rpart. Additionally, by comparing the overall accuracy (OA) and kappa coefficients of various models across 100 repeated trials, we observe that the SSC-SL, SSC-Ranger, SL, and Ranger models demonstrate relatively lower uncertainty. In contrast, the SSC-Rpart, SSC-XGBoost, Rpart, and XGBoost models exhibit comparatively higher uncertainty. This suggests that the Ranger model is relatively robust. Similarly, the SL model, which integrates three basic learners including Ranger, is also exceptionally robust. In summary, the SSC-SL model combines multiple base learners and effectively utilises a significant amount of unlabelled data, addressing the issue of low predictive accuracy and limited soil sample data in the

digital soil mapping process. It can serve as an effective method to enhance the accuracy of digital soil mapping predictions.

According to the spatial distribution maps of soil subgroups, we can see that the overall distribution range of Fe-accumuli-Stagnic Anthrosols (Typic Fe-accumuli-Stagnic Anthrosols and Endogleyic Fe-accumuli-Stagnic Anthrosols) is essentially consistent across all models. These soils are typically located in paddy field areas, whereas Typic Hapli-Udic Argosols are found in dry or forest lands. This underscores the necessity of remote sensing data, particularly land use data, in determining the distribution ranges of these soil types. Similarly, remote sensing data, in conjunction with topographic data, plays a crucial role in distinguishing Typic Hapli-Udic Cambosols from other soil types. For instance, Typic Hapli-Udic Cambosols are mainly distributed in composite areas of arbour forest land and the mid-slopes of Gaoli Mountain. Additionally, soil parent material affects the distribution of Red Ferri-Udic Argosols, Red Ferri-Udic Cambosols, Typic Purpli-Udic Cambosols, and other soils. Notably, Red Ferri-Udic Argosols and Red Ferri-Udic Cambosols are influenced by both soil parent material and topographic location. Topographic variation is also significant in differentiating Typic Fe-accumuli-Stagnic Anthrosols and Endogleyic Fe-accumuli-Stagnic Anthrosols. This indicates that soil types are influenced or characterised by multiple factors, with remote sensing data, soil parent material, and topographic elements all playing key roles. It is worth noting that the Rpart model showed the lowest prediction accuracy for Typic Udi-Orthic Primosols, failing to adequately depict their spatial distribution, unlike other models. This may be due to the stony soils being primarily influenced by topographical factors, predominantly located at the top of Gaoli Mountain. The Rpart model, being relatively simplistic, struggles to depict the spatial distribution of soil types formed under these extreme conditions with the existing environmental variables. In contrast, all models showed high prediction accuracy for Typic Purpli-Udic Cambosols, as their distribution within the study area is solely influenced by soil parent material, which is well represented in existing soil parent material maps. Additionally, the prediction accuracy for Fe-accumuli-Stagnic Anthrosols (Typic Fe-accumuli-Stagnic Anthrosols and Endogleyic Fe-accumuli-Stagnic Anthrosols) was also high due to their distribution in paddy fields being well represented in current land use maps, highlighting the potential of remote sensing data in effectively identifying certain soil types.

5. Conclusions

This study successfully developed and evaluated the SSC-SL model, integrating SSC with SL using foundational learners Ranger, Rpart, and XGBoost, combined with FCM for selecting unlabelled samples. The results indicate the following:

1. The SSC-SL model exhibits robust performance in soil classification mapping, significantly enhancing prediction accuracy. By incorporating unlabelled sample points through the SSC approach and leveraging the compounded strengths of individual models in the SL framework, the SSC-SL model effectively addresses the limitations of the prediction degradation caused by the selection of a simple machine learning model.
2. The soil subgroup spatial distribution maps generated by the SSC-SL model are more rational, improving upon the spatial distribution ranges of different soil types in the foundational models. This model also offers a clearer depiction of details.
3. SPM, LU, MRVBF, and Ele consistently rank as highly important in all eight models, indicating that they are the primary environmental variables influencing the soil types in the study area.

Author Contributions: Conceptualisation, F.Z. and J.P.; methodology, F.Z.; software, F.Z.; investigation, F.Z., C.Z., W.L. and Z.F.; data curation, F.Z. and C.Z.; writing—original draft preparation, F.Z., Z.L. and J.P.; visualisation, F.Z. and W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number: 41971057.

Data Availability Statement: Data are contained within the article.

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

References

1. Wagg, C.; Bender, S.F.; Widmer, F.; Van Der Heijden, M.G.A. Soil Biodiversity and Soil Community Composition Determine Ecosystem Multifunctionality. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 5266–5270. [CrossRef]
2. Amundson, R.; Berhe, A.A.; Hopmans, J.W.; Olson, C.; Sztein, A.E.; Sparks, D.L. Soil and Human Security in the 21st Century. *Science* **2015**, *348*, 1261071. [CrossRef]
3. Ippolito, T.A.; Herrick, J.E.; Dossa, E.L.; Garba, M.; Ouattara, M.; Singh, U.; Stewart, Z.P.; Prasad, P.V.V.; Oumarou, I.A.; Neff, J.C. A Comparison of Approaches to Regional Land-Use Capability Analysis for Agricultural Land-Planning. *Land* **2021**, *10*, 458. [CrossRef]
4. Alhajj Ali, S.; Vivaldi, G.A.; Garofalo, S.P.; Costanza, L.; Camposeo, S. Land Suitability Analysis of Six Fruit Tree Species Immune/Resistant to Xylella Fastidiosa as Alternative Crops in Infected Olive-Growing Areas. *Agronomy* **2023**, *13*, 547. [CrossRef]
5. Poggio, L.; De Sousa, L.M.; Batjes, N.H.; Heuvelink, G.B.M.; Kempen, B.; Ribeiro, E.; Rossiter, D. SoilGrids 2.0: Producing Soil Information for the Globe with Quantified Spatial Uncertainty. *Soil* **2021**, *7*, 217–240. [CrossRef]
6. Liu, F.; Wu, H.; Zhao, Y.; Li, D.; Yang, J.-L.; Song, X.; Shi, Z.; Zhu, A.-X.; Zhang, G.-L. Mapping High Resolution National Soil Information Grids of China. *Sci. Bull.* **2022**, *67*, 328–340. [CrossRef]
7. Žížala, D.; Minařík, R.; Skála, J.; Beitlerová, H.; Juřicová, A.; Reyes Rojas, J.; Penížek, V.; Zádorová, T. High-Resolution Agriculture Soil Property Maps from Digital Soil Mapping Methods, Czech Republic. *Catena* **2022**, *212*, 106024. [CrossRef]
8. Lembrechts, J.J.; Ashcroft, M.B.; Frenne, P.D.; Kemppinen, J.; Kopecký, M.; Luoto, M.; Maclean, I.M.D.; Crowther, T.W.; Bailey, J.J.; Haesen, S.; et al. Global Maps of Soil Temperature. *Glob. Chang. Biol.* **2022**, *28*, 3110–3144. [CrossRef]
9. Ivushkin, K.; Bartholomeus, H.; Bregt, A.K.; Pulatov, A.; Kempen, B.; De Sousa, L. Global Mapping of Soil Salinity Change. *Remote Sens. Environ.* **2019**, *231*, 111260. [CrossRef]
10. Asgari, N.; Ayoubi, S.; Jafari, A.; Demattê, J.A.M. Incorporating Environmental Variables, Remote and Proximal Sensing Data for Digital Soil Mapping of USDA Soil Great Groups. *Int. J. Remote Sens.* **2020**, *41*, 7624–7648. [CrossRef]
11. Teng, H.; Viscarra Rossel, R.A.; Shi, Z.; Behrens, T. Updating a National Soil Classification with Spectroscopic Predictions and Digital Soil Mapping. *Catena* **2018**, *164*, 125–134. [CrossRef]
12. Cao, D.; Xing, H.; Wong, M.S.; Kwan, M.-P.; Xing, H.; Meng, Y. A Stacking Ensemble Deep Learning Model for Building Extraction from Remote Sensing Images. *Remote Sens.* **2021**, *13*, 3898. [CrossRef]
13. Cui, S.; Yin, Y.; Wang, D.; Li, Z.; Wang, Y. A Stacking-Based Ensemble Learning Method for Earthquake Casualty Prediction. *Appl. Soft Comput. J.* **2021**, *101*, 107038. [CrossRef]
14. Faska, Z.; Khrissi, L.; Haddouch, K.; El Akkad, N. A Robust and Consistent Stack Generalized Ensemble-Learning Framework for Image Segmentation. *J. Eng. Appl. Sci.* **2023**, *70*, 74. [CrossRef]
15. Aydın, Y.; Işıkdağ, Ü.; Bekdaş, G.; Nigdeli, S.M.; Geem, Z.W. Use of Machine Learning Techniques in Soil Classification. *Sustainability* **2023**, *15*, 2374. [CrossRef]
16. Sharififar, A.; Sarmadian, F.; Malone, B.P.; Minasny, B. Addressing the Issue of Digital Mapping of Soil Classes with Imbalanced Class Observations. *Geoderma* **2019**, *350*, 84–92. [CrossRef]
17. van Engelen, J.E.; Hoos, H.H. A Survey on Semi-Supervised Learning. *Mach. Learn.* **2020**, *109*, 373–440. [CrossRef]
18. Kostopoulos, G.; Karlos, S.; Kotsiantis, S.; Ragos, O. Semi-Supervised Regression: A Recent Review. *IFS* **2018**, *35*, 1483–1500. [CrossRef]
19. Taghizadeh-Mehrjardi, R.; Schmidt, K.; Amirian-Chakan, A.; Rentschler, T.; Zeraatpisheh, M.; Sarmadian, F.; Valavi, R.; Davatgar, N.; Behrens, T.; Scholten, T. Improving the Spatial Prediction of Soil Organic Carbon Content in Two Contrasting Climatic Regions by Stacking Machine Learning Models and Rescanning Covariate Space. *Remote Sens.* **2020**, *12*, 1095. [CrossRef]
20. Zhang, L.; Yang, L.; Ma, T.; Shen, F.; Cai, Y.; Zhou, C. A Self-Training Semi-Supervised Machine Learning Method for Predictive Mapping of Soil Classes with Limited Sample Data. *Geoderma* **2021**, *384*, 114809. [CrossRef]
21. Zhu, C.; Wei, Y.; Zhu, F.; Lu, W.; Fang, Z.; Li, Z.; Pan, J. Digital Mapping of Soil Organic Carbon Based on Machine Learning and Regression Kriging. *Sensors* **2022**, *22*, 8997. [CrossRef]
22. Fang, Z.; Lu, W.; Zhu, F.; Zhu, C.; Li, Z.; Pan, J. Landscape Classification System Based on RKM Clustering for Soil Survey UAV Images—Case Study of the Small Hilly Areas in Jurong City. *Sensors* **2022**, *22*, 9895. [CrossRef]
23. Chinese Soil Taxonomy Research Group. *Keys to Chinese Soil Taxonomy*, 3rd ed.; University of Science and Technology of China Press: Hefei, China, 2001.
24. Jenny, H. *Factors of Soil Formation: A System of Quantitative Pedology*; McGraw-Hill: New York, NY, USA, 1941.
25. Demattê, J.A.M.; da Silva Terra, F. Spectral Pedology: A New Perspective on Evaluation of Soils along Pedogenetic Alterations. *Geoderma* **2014**, *217–218*, 190–200. [CrossRef]
26. Li, Y.; Zhao, Z.; Wei, S.; Sun, D.; Yang, Q.; Ding, X. Prediction of Regional Forest Soil Nutrients Based on Gaofen-1 Remote Sensing Data. *Forests* **2021**, *12*, 1430. [CrossRef]

27. Marchetti, A.; Piccini, C.; Santucci, S.; Chiuchiarelli, I.; Francaviglia, R. Simulation of Soil Types in Teramo Province (Central Italy) with Terrain Parameters and Remote Sensing Data. *Catena* **2011**, *85*, 267–273. [CrossRef]
28. Zeraatpisheh, M.; Ayoubi, S.; Jafari, A.; Finke, P. Comparing the Efficiency of Digital and Conventional Soil Mapping to Predict Soil Types in a Semi-Arid Region in Iran. *Geomorphology* **2017**, *285*, 186–204. [CrossRef]
29. Wilson, M.J. The Importance of Parent Material in Soil Classification: A Review in a Historical Context. *Catena* **2019**, *182*, 10413. [CrossRef]
30. Wadoux, A.M.J.-C. Using Deep Learning for Multivariate Mapping of Soil with Quantified Uncertainty. *Geoderma* **2019**, *351*, 59–70. [CrossRef]
31. Kuhn, M. caret: Classification and Regression Training. R Package Version 6.0-92. Available online: https://CRAN.R-project.org/package=caret (accessed on 2 December 2023).
32. Ganaie, M.A.; Hu, M.; Malik, A.K.; Tanveer, M.; Suganthan, P.N. Ensemble Deep Learning: A Review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151. [CrossRef]
33. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
34. Wright, M.N.; Ziegler, A. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in *C++* and *R. J. Stat. Soft.* **2017**, *77*, 1–17. [CrossRef]
35. Breiman, L. *Classification and Regression Trees*; Chapman & Hall: New York, NY, USA, 1984.
36. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
37. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2023; Available online: https://www.R-project.org/ (accessed on 2 December 2023).
38. RStudio Team. *RStudio: Integrated Development for R*; RStudio, PBC: Boston, MA, USA, 2020. Available online: http://www.rstudio.com/ (accessed on 2 December 2023).
39. Coyle, J.; Hejazi, N.; Malenica, I.; Phillips, R.; Sofrygin, O. sl3: Pipelines for Machine Learning and Super Learning, R Package Version 1.4.4. Available online: https://github.com/tlverse/sl3 (accessed on 2 December 2023).
40. Van Der Laan, M.J.; Polley, E.C.; Hubbard, A.E. Super Learner. *Stat. Appl. Genet. Mol. Biol.* **2007**, *6*. [CrossRef]
41. Dunn, J.C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *J. Cybern.* **1973**, *3*, 32–57. [CrossRef]
42. Li, Y.; Zhang, C.; Li, Z.; Yang, L.; Jin, X.; Gao, X. Analysis on the Temporal and Spatial Characteristics of the Shallow Soil Temperature of the Qinghai-Tibet Plateau. *Sci. Rep.* **2022**, *12*, 19746. [CrossRef]
43. Peng, Y.; Roell, Y.E.; Odgers, N.P.; Møller, A.B.; Beucher, A.; Greve, M.B.; Greve, M.H. Mapping and Describing Natural Terroir Units in Denmark. *Geoderma* **2021**, *394*, 115014. [CrossRef]
44. Dunkl, I.; Ließ, M. On the Benefits of Clustering Approaches in Digital Soil Mapping: An Application Example Concerning Soil Texture Regionalization. *Soil* **2022**, *8*, 541–558. [CrossRef]
45. Gelb, J.; Apparicio, P. Apport de la classification floue c-means spatiale en géographie: Essai de taxinomie socio-résidentielle et environnementale à Lyon. *Cybergeo* **2021**, *972*, 1–26. [CrossRef]
46. Estévez, V.; Beucher, A.; Mattbäck, S.; Boman, A.; Auri, J.; Björk, K.-M.; Österholm, P. Machine Learning Techniques for Acid Sulfate Soil Mapping in Southeastern Finland. *Geoderma* **2022**, *406*, 115446. [CrossRef]
47. Yang, X.; Song, Z.; King, I.; Xu, Z. A Survey on Deep Semi-Supervised Learning. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 8934–8954. [CrossRef]
48. Palomares Alabarce, F.J. SSLR: Semi-Supervised Classification, Regression and Clustering Methods, R Package Version 0.9.3.3. Available online: https://CRAN.R-project.org/package=SSLR (accessed on 2 December 2023).
49. Tao, S.; Zhang, X.; Feng, R.; Qi, W.; Wang, Y.; Shrestha, B. Retrieving Soil Moisture from Grape Growing Areas Using Multi-Feature and Stacking-Based Ensemble Learning Modeling. *Comput. Electron. Agric.* **2023**, *204*, 107537. [CrossRef]
50. Cahyana, D.; Barus, B.; Darmawan; Mulyanto, B.; Sulaeman, Y. Assessing Machine Learning Techniques for Detailing Soil Map in the Semiarid Tropical Region. *IOP Conf. Ser. Earth Environ. Sci.* **2021**, *648*, 012018. [CrossRef]
51. Assami, T.; Hamdi-Aïssa, B. Digital Mapping of Soil Classes in Algeria—A Comparison of Methods. *Geoderma Reg.* **2019**, *16*, e00215. [CrossRef]