Article

Classification And Identification of Organic Matter in Black Soil Based on Simulated Annealing Optimization of LSVM-Stacking Model

Zifang Zhang [1], Zhihua Liu [2], Hongzhao Xu [1], Qinghe Zhao [1], Junlong Fang [1]* and Kezhu Tan [1]*

[1] Electrical Engineering and Information College, Northeast Agricultural University, Harbin, China; zhangzifang@neau.edu.cn (Z.Z.); xuhongzhao9858@neau.edu.cn (H.X.); zhaoqinghe@neau.edu.cn (Q.Z.);
[2] Resources and Environment Collage, Northeast Agricultural University, Harbin, China; zhihua-liu@neau.edu.cn (Z.L.);
* Correspondence: jlfang@neau.edu.cn (J.F.); Tel.: +86-189-4505-5858; kztan@neau.edu.cn (K.T.); Tel.: +86-451-5519-0446

Abstract: For the soil in different regions, the nutrient fertility contained in it is different, and the detection and zoning management of soil nutrients before tillage every year can improve grain yield. In this paper, an integrated learning strategy model based on black soil hyperspectral data is designed for rapid classification of organic matter content classification of black soil. Soil hyperspectral image dataset of Xiangyang Experimental Base was collected; by changing the internal structure of the stacking model, an LSVM-stacking model with (MLP, SVC, DTree, XGBl, kNN) five classifiers as the L1 layer was built, and the simulated annealing algorithm was used for hyperparameter optimization. Compared to other stacking models, the LSVM-stacking metrics are significantly improved. The accuracy rate of hyperparameter optimization is improved by 38.6515%, the accuracy rate of the independent test data set is 0.9488, and the comparison of individual learners can improve the recognition classification accuracy of label"1" to 1.0.

Keywords: Hyperspectral Technology; Non-destructive Testing; Black Soil; Ensemble learning; Support Vector Machine

1. Introduction

Cultivated land is an essential non-renewable natural resource in agricultural production and a natural complex with many components. Soil organic matter (SOM) content is an essential indicator of soil fertility and is directly related to crop growth and yield[1]. The black soil area of Northeast China is one of the three world-famous black soil areas. Because it is rich in humus and has high soil fertility, it is a kind of soil that is very suitable for the growth of crops and is a valuable cultivated land resource in China.

In the second national soil census, the soil was divided into six grades according to the content of SOM.:<0.6 g/kg,0.6 g/kg-10 g/kg,10 g/kg-20 g/kg,20 g/kg-30 g/kg,30 g/kg-40 g/kg,40 g/kg<. According to the SOM content, the land is managed in the grid and cultivated in different zones. Every year, before spring plowing, a comprehensive assessment of the black soil area is carried out to guide farmers to cultivate scientifically, ensuring the people’s livelihood and contributing to agriculture’s sustainable development. The routine detection method of soil organic matter requires field sampling and laboratory chemical sample analysis, followed by geographic interpolation to map the spatial distribution.

In Feb. 2022, the General Office of the State Council, PRC, issued a notification to decide to complete the third national soil census from 2022 to 2025[2]. It is time-consuming and laborious to conduct large-scale field surveys and laboratory sample analyses and draw spatial maps. Agricultural informatization is the upsurge of global agricultural development in the 21st century. Spectroscopy and imaging technology have been applied

doi:10.20944/preprints202209.0239.v1

in agricultural production by multi-scale agricultural remote sensing platforms such as satellites and UAV systems. Hyperspectral technology is widely used in multi-objective classification, such as agricultural product classification[3–12], the detection of nutrients in agricultural products[13–15], and the determination of SOM[16–21]. Compared with traditional chemical determination methods, hyperspectral is fast and non-destructive. Reis et al. [22] collected from 8 depths near the COAMO in southern Brazil, and the AsiaFENIX sensor was used to collect spectral images in the 380-2506 nm band and established a SOM content estimation model by PLSR. Rapid monitoring of SOM content to achieve rapid grading of black soil fertility has particular guiding significance for adapting measures to local conditions, scientific farming, and sustainable agricultural development.

Any single classifier has its advantages and disadvantages, and ensemble learning can take advantage of some combination rules to improve the model. Ensemble learning can comprehensively use each learner's advantages, integrate the model's prediction results, and make up for the shortcomings of a single model that is greatly disturbed by sensitive samples and lacks robustness[23]; therefore, it is also the focus of current research. Many previous studies have demonstrated that the ensemble learning method has a good effect on hyperspectral classification[24–30]. Ensemble learners based on SVM perform well on high-dimensional hyperspectral data[24–26]. The classification accuracy using the decision tree is similar to that of SVM, but the decision tree model is susceptible to noise[27]. Random Forest is also a standard classification algorithm, but it is computationally expensive compared to other algorithms[28–30]. The standard ensemble methods are homogeneous ensembles, Guo et al.[31] proposed a heterogeneous ensemble model that integrates SVM, KELM, and MLR by bagging, which also has good applicability. Both theoretically and empirically, the integration model is not as big as possible. Conversely, sometimes small ensemble models may have unexpected effects, so the selection of classifiers and the way of the ensemble are issues that need to be solved urgently.

In this paper, an ensemble learning strategy model based on black soil hyperspectral data is designed to complete the gradation of SOM content. The LSVM-stacking model was screened out by comparing nine stacking models, and simulated annealing was used for hyperparameters optimization. The accuracy score on the test set is 0.6318 before the parameter adjustment and 0.8760 after the parameter adjustment. In the independent validation dataset, the LSVM-stacking model has a higher accuracy of 0.9488 than other stacking models, and the recognition classification accuracy of label"1" can be improved to 1.0. It means that the LSVM-stacking model has the best applicability.

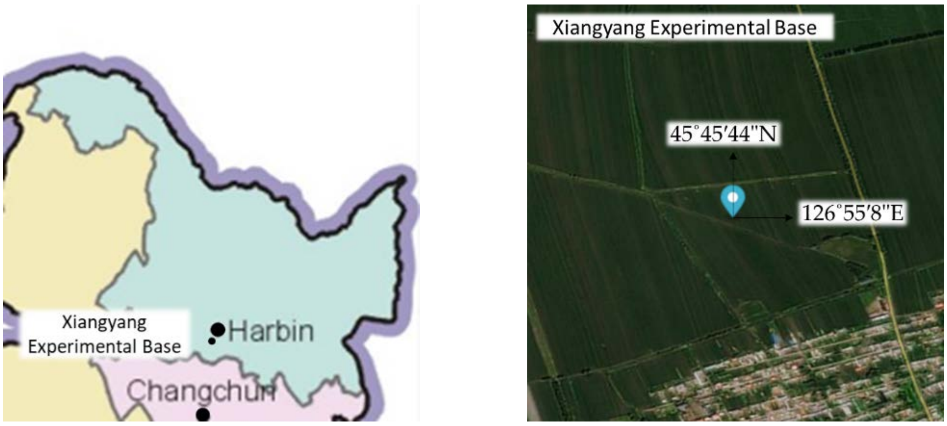
2. Materials and Methods

2.1 Materails

2.1.1 Determination of SOM

Soil samples were collected from Xiangyang Experimental Base of Northeast Agricultural University (45°45′44"N, 126°55′8"E), where the northeast China, near Harbin, Heilongjiang Province. It has a temperate monsoon climate, with long, cold, and dry winters, short·, hot, and rainy summers, and average annual precipitation of 569.1 mm. Figure 1 shows the geographical location of the Xiangyang Experimental Base.

doi:10.20944/preprints202209.0239.v1



(a) Location of the Xiangyang Experimental Base.

(b) The geographical coordinates and the geographical environment of the Xiangyang Experimental Base.

Figure 1. The geographical location of the Xiangyang Experimental Base

When sampling, drilling holes with a post-hole digger and installing a well casing to prevent the collapse of the well wall during sampling affect the accuracy of the sample. Soil samples at four depths of 0-10cm, 10-20cm, 20-30cm, and 30-40cm were taken by five-point mixed sampling. A total of 112 soil samples were collected, and the location of the sampling point was recorded with a handheld GPS at the same time. After returning it to the laboratory, the soil samples taken from each sampling point are mixed evenly, and impurities such as rhizomes are removed. After natural drying, take a part of the ground soil and pass it through a 20-mesh sieve, and this part of the sample is used to measure the spectral curve. The other part was ground and passed through a 100-mesh sieve, and this part of the sample was used for chemical analysis.

SOM data were provided by the College of Resources and Environment, Northeast Agricultural University. The analysis of SOM content was done by the potassium dichromate method[32]. Soil mixed with excess potassium dichromate solution, the heat outside the oil bath, potassium dichromate can oxidize organic matter in the soil, and titrate the remaining potassium dichromate with a standard solution of ferrous iron, then calculate the content of organic matter in the sample from the consumption of potassium dichromate.

Since the sample has been dried and ground in powder form, placing the samples directly on the working platform for testing will cause the work platform to malfunction. Therefore, select a petri dish with a diameter of 6 cm and a depth of 1 cm, and the samples are loaded into the petri dish and labeled it. Flatten the sample to avoid shadows caused by halogen lamps. Pick out small grassroots to avoid interference in the subsequent data processing.

Spectral data acquisition was performed using the Headwall's Hyperspec® VNIR family of integrated hyperspectral imaging A-sensors. The sample collection needs to be carried out in a dark room, the light source should be limited to the halogen light on the experimental platform as much as possible, and calibrating sensors with a standard whiteboard before each sample. The sensor wavelength range is configured from 400nm to 1000nm, the spectral resolution is about 3nm, and 203 bands are collected. During the acquisition, the moving speed of the push-broom platform was 5mm/s, the exposure time was 38.84ms, the frame period was 0.04ms, and the final resolution of the obtained cubic image was (1004, 812, 203).

Completing radiometric correction with the white and dark reference files in ENVI Clasic5.3. Use the ROI Type square function on the processed image to create Regions of

Interest (ROI) on the sample. To avoid reflections from the edges of the petri dish, only take the center part. Each ROI displays an average of about 240 pixels, and the software calculates the spectral average of the pixels in the region to obtain the spectral curve of this ROI.

The 3063 sample points are divided into two groups according to the ratio of 7:3; the training set and testing set are obtained. Another 100 groups of samples were selected to extract 2831 sample points as the validation set. The average spectral curves of the three types are shown in Figure 2. The range of organic matter content in various samples and the specific quantity of samples are shown in Table 1.

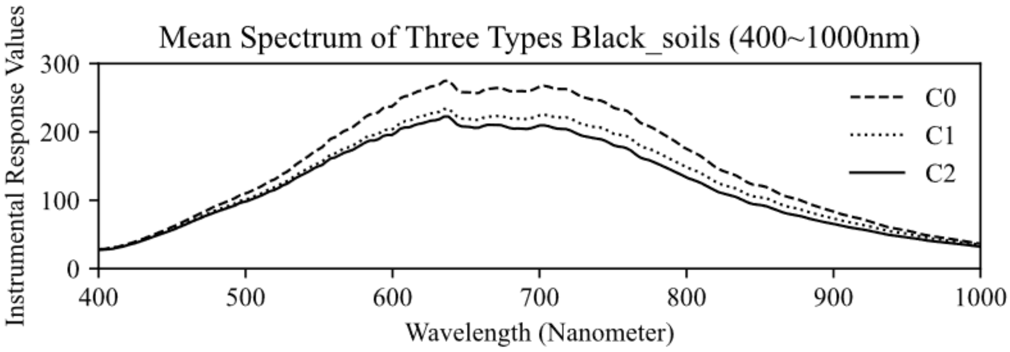


Figure 2. Mean spectral curves of three-grade black soils.

Table 1. The range of organic matter content in three-grade black soil samples and the specific quantity of sample.

	organic matter (mg/kg)	**Training set**	**Testing set**	**total**	**validation set**	**total**
type 0	10.00-19.99	709	304	1013	704	1717
type 1	20.00-29.99	721	309	1030	1069	2099
type 2	30.00-45.00	714	306	1020	1058	2078
total		2144	919	3063	2831	5894

2.2 Methods

2.2.1 Ensemble Learning

For a multi-classifier system, only including individual learners of the same type is called a homogeneous ensemble, while the individual learners in a heterogeneous ensemble are generated by different types of algorithms[23].

The combination strategy of ensemble learning refers to the cooperation strategy between individual learners. Standard ensemble methods include boosting, bagging, and stacking. Like AdaBoost and Gradient Boosting, Boosting integrates a series of the individual learner, like Decision Tree, which realizes homogenous integration of serial training through residuals or other indicators for reducing bias. Bagging integrates bootstrap sampling to form a sampled sub-set in this process for training individual learners. The trained homogeneous learners are trained in parallel to reduce variance. Stacking is a broader ensemble strategy, hierarchically stacking heterogeneous individual learners; the statistical result of one-layer individual learners is input as features to two-layer meta-learners to complete the stacking algorithm. Variance is reduced by parallel training, and deviation is reduced by serial training.

No matter which strategy is employed, ensemble learning is expected to obtain multiple aspects of the sample space through multiple angles with significant differences and achieve better-supervised learning effects than individual learners.

The stacking strategy can more flexibly select different heterogeneous learners for integration. The stacking strategy is divided into two layers: the L1 layer is composed of several individual learners, and the strategic decision layer consists of meta-learner L2.

Each learner in the L1 will complete supervised learning in the sample space, convert the original data into n transition data S, and then input it into the L2. The most basic form of the L2 layer is equal voting (classification problem) or equal weighting (regression problem), but a strong learner is generally adopted as the core of this layer.

The selection of the L1 individual learner needs to consider the feature information in the sample space because the individual learners of this layer directly perform feature extraction and decision output on the samples. So, the selection needs to consider general feature engineering problems, such as the linear model is possibly a strong individual learner in the high-dimensional small sample data set. However, when the features are highly collinear, the linear model may become a weak individual learner. The sample subspace input to the L2 is quite different from the original space, and the choice of the learner is more inclined to the features of S. In the classification problem, the output of the L1 can be sample labels or the continuous cross probability or information entropy of the sample labels. The most significant difference between the stacking， boosting, and bagging strategies is the existence of the L2. Stacking does not target a single individual learner for further learning. When the sample space is determined, after using multiple and different individual learners to build the L1, determining a reasonable strong learner in the L2 layer will significantly improve the learning effect of the model. The pseudo-code of stacking is as follows:

Table 2. The pseudo-code of stacking algorithms.

```
Algorithm. Ensemble Learning (for Stacking)
Input: training set: $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$;
       individual learners: $\eta_1, \eta_1, \dots, \eta_T$;
       meta-learners: $\eta$.
for $t = 1, 2, \dots, T$ do
$h_t = \eta_t(D)$;
end for
$S = \emptyset$;
for $i = 1, 2, \dots, m$ do
  for $t = 1, 2, \dots, T$ do
    $s_{it} = h_t(x_i)$;
  end for
  $S = ((s_{i1}, s_{i2}, \dots, s_{iT}), y_i)$;
end for
$h' = \eta(D')$;
Output: $H(x) = h'(h_1(x), h_2(x), \dots, h_T(x))$
```

where D is a training set with m samples, L1 consists of T individual learners. L2 is meta-learner η. Receiving trained model h_t with the training set. For x_i in D, $s_{it} = \mathrm{h}_t(x_i)$, the second training set generated by x_i is $s_i = (s_{i1}, s_{i2}, \dots, s_{iT})$, the label is y_i. The secondary training set produced by T individual learners is $S = \{(s_i, y_i)\}_{i=1}^{m}$, which will be used for training meta-learner.

In conclusion, when choosing the stacking for ensemble learning model construction, the diversity of L1 individual learners and the selection of L2 meta-learners are the two decisive aspects of this strategy.

2.2.2 Individual Learners Selection

The following five basic models are selected as the individual learner of the L1:

- CART Tree (DTree)

The Decision Tree algorithm is representative of the tree model. The Decision Tree algorithm is a binary growing tree, a nonlinear model that can realize data visualization. The time cost of model training is $O(n * m^2 * \log m)$. This model uses the Gini index as the segmentation index when measuring tree growth. In this individual learner, the minimum number of samples that model leaf nodes can split is 2, and the minimum number of samples for leaf nodes is 1. However, the maximum depth of model splitting and the maximum leaf node is not limited to enable the model to be adequately fitted by a greedier algorithm.

- RBF-SVC (SVM)

Support Vector Machines implemented with Radial Basis Function Kernel SVM. The lower bound of the model training time cost is $O(n*m^2)$, and the upper bound is $O(n*m^3)$. In this individual learner, the regularization parameter C is constrained to be 1, and the kernel coefficient gamma is 0.0016.

- k-nearest neighbor（kNN)

The k-nearest neighbor model is nonlinear and is often used in problems with unclear classification boundaries. The lower bound of the model training time cost is $O(n*\log(m))$, and the upper bound is $O(n*m)$. The nearest neighbor calculation algorithm is k-DTree. The number of neighbors that can be queried is 5. All neighbors in the neighborhood have the same weight, that is, uniform weights. The spatial distance is calculated by Euclidean distance.

- Multilayer Perceptron (MLP3)

The multilayer perceptron is a fully connected feedforward network model. This time, the three-layer perceptron model is a relatively simple perceptron model with only one hidden layer. The number of nodes in each model layer is (100, 50, 25). Activation function for the hidden layer is "relu", that is, $f(x) = \max(0, x)$. The solver for weight optimization is adam, that is, a stochastic gradient-based optimization proposed by Kingma, Diederik, and Jimmy Ba. Moreover, the maximum number of iterations is 200. The time cost of training this model is $O(n*m*hiddens^{neurons}*outputs*inters)$.

- XGBoost (XGBl)

The XGBoost algorithm integrates n linear models as the representative of the linear model. Set the booster parameter to "gblinear" to integrate linear models, and there are also no restrictions on the maximum depth and maximum leaf nodes for model splits.

In the time complexity representation, m represents the sample size, and n represents the sample dimension.

2.2.3 L2 selection

In this experiment, individual learners are determined as the L1 through a heterogeneous ensemble strategy to consider various types of classifiers that reduce the model's variance. The algorithm is completed by matching different learners as the L2, and the structure is shown in Figure 3. The L1 is set to five individual learners in 3.1. After converting the original data into transition data, the L2 is connected to complete the classification. The L2 function candidate models are Logist, MLP, kNN, XGBoost, Decision Tree, LSVM, SVC, Random Forest, and AdaBoost.

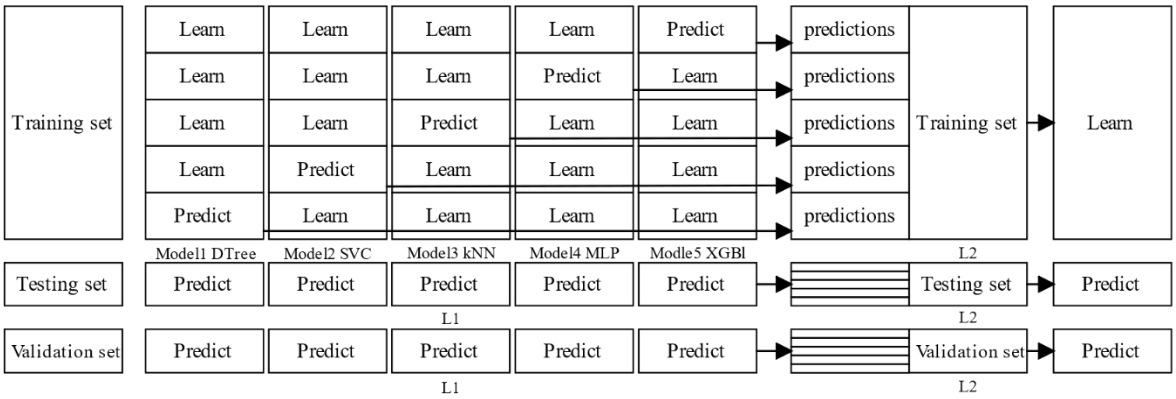


Figure 3. The process structure of the algorithm.

This stacking method has a process structure similar to multilayer neural networks. This paper mainly conducts comparative experiments on model selection at the L2. When selecting the L2 function, it is necessary to adjust the parameters of the objective function. The simulated annealing hyperparameter optimization algorithm is used for parameter configuration, and the optimal parameter solution is selected after 1000 iterations. By comparing the trained models on the testing set, the optimal L2 function is selected. In addition, an independent validation set is introduced to verify the model's applicability.

2.2.4 Simulated Annealing

Hyperparameters are parameters used to control the behavior of an algorithm when building a model, which cannot be obtained from regular training and must be set manually. One of the most complex parts of machine learning is finding the best hyperparameters for the model. The performance of the model is directly related to the hyperparameters. The better the hyperparameters are tuned, the better the model's performance.

The simulated annealing algorithm was first proposed by N. Metropolis et al. [33]. However, its use in combinatorial optimization design was proposed in 1983 by S. Kirkpatrick et al. and V. Cerny [34–36]. It is derived from the principle of metal quench cooling. When the metal heats up and melts, the thermal energy is converted into kinetic energy, and the particles in the metal start to move disorderly. When the metal cools down slowly, the particles tend to be ordered; when the metal cools down to a normal temperature state, the kinetic energy is the lowest[37]. The simulated annealing algorithm consists of the Metropolis criterion and the annealing process. The annealing process is understood as finding the optimal global solution. Moreover, the purpose of the Metropolis criterion is to search for the optimal global solution out of the optimal local solution, which is the basis for annealing.

Metropolis criterion is generally expressed as follows (1):

$$P = \begin{cases} 1, & E(x_{new}) < E(x_{old}) \\ \exp\left(-\frac{E(x_{new}) - E(x_{old})}{T}\right), & E(x_{new}) \geq E(x_{old}) \end{cases} \tag{1}$$

The Metropolis criterion states that at temperature T, there is a probability $P(\Delta E)$ of cooling with an energy difference ΔE, expressed as $P(\Delta E) = exp(\Delta E/(kT))$, where k is the Boltzmann constant, exp is the natural exponent, and $\varDelta E < 0$. So P and T are positively correlated. This formula means that the higher the temperature, the greater the probability of cooling with an energy difference of ΔE; the lower the temperature, the lower probability. If the energy attenuation, then this change will be accepted with 1. If the energy does not change or increase, this change deviates from the direction of the optimal global solution, which will be accepted with P. Because the temperature gradually decreases during the annealing process, $\varDelta E$ is always less than 0. Therefore, $\Delta E/kT < 0$, so the range of $P(\Delta E)$ is $(0,1)$. With the decrease of temperature T, $P(\Delta E)$ will gradually decrease and eventually stabilize to achieve the optimal global solution.

2.2.5 Evaluation indicators

(1) Accuracy and Class Accuracy

For classifying organic matter content in black soil, which is essentially a multi-classification problem in supervised learning, accuracy (ACC) can be used as the model evaluation index. Accuracy is calculated as follows (2):

$$acc(y, y_{pred}) = \frac{1}{N}\sum_{i=0}^{N} I\left(y_{pred_i} = y_i\right) \tag{2}$$

where y_i is the true label of the i sample, y_pred_i is the predicted label of sample i, and N is the total number of samples.

Class accuracy (C-ACC) for evaluating single class classification is a variant of accuracy, which indicates the proportion of a that the model predicts correctly in the label. The formula is as follows (3):

$$acc^j(y, y_{pred}) = \frac{1}{N^j}\sum_{i=0}^{N^j} I\left(y_{pred_i} = y_i\right), y_i \in label(j) \tag{3}$$

where y_i is the true label of the i sample, y_pred_i is the predicted label of sample i, and N^j is the total number of samples of this label.

The legal range of both of them is [0, 1]. The closer it is to 1, the higher the proof accuracy and the better the classification effect of the model. This study balances the number of samples in the three labels. So, for this three-label problem, the lowest limit of the total accuracy rate should be 0.33.

(2) F score

F score, also known as the balance score, is the weighted average of precision and Recall. In this classification problem, it is necessary to consider the precision and the recall; that is, the F1 score (F1) is quoted. The formula is as follows (4):

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

where precision and recall respectively represent the precision and recall within the label, and the formulas are as follows (5) and (6):

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$re\ call = \frac{TP}{TP + FN} \tag{6}$$

where TP represents the number of correctly predicted samples, FP represents the number of wrongly predicted samples from other classes as this class, and FN represents the number of samples from this class that is incorrectly predicted as other classes.

The legal range of F score is [0, 1], an enormous value means a better model.

2.2.5 Compute environment

All the codes designed in this research have been open-sourced under the MIT license. The hardware environment that the analysis depends on is shown in Table 3 below, and the compile environment is based on Python 3.9 in Windows 10 LTSC. To ensure reproducibility of all experimental and analysis results, all random seeds involved in this paper are set as 615.

Table 3. Environment and tools of analysis and model building in paper.

Computing Environment		**Algorithmic Environment**
CPU	Intel® Core™ i5-10400 (2.90GHz)	Scikit-learn 1.0.1,
GPU	Nvidia GeForce RTX 3070	Numpy 1.18.5,
RAM	DDR4 3000Mhz 16GB = 2x8GB	Pandas 1.3.3,
operating system	Windows LTSC 21H2	Xgboost 1.4.2,
Random seed	615	Scipy 1.5.0

3. Results

3.1. stacking model building

The independent test results of the five individual learners in the L1 layer are as follows in Figure 4. Among the five classification models, kNN and XGBl have higher ACC in the test set, which can reach 0.8857 and 0.8520, indicating that these two individual learners have excellent essential performance and can be initially designated as solid classifiers. Compared with the high accuracy index, the F1 of kNN is 0.8859, which further illustrates the strong classifier properties of kNN in this sample space. The ACC of DTree, SVM, and MLP is relatively low, which are 0.8313, 0.8313, and 0.8292, respectively, so they will be initially defined as weak classifiers.

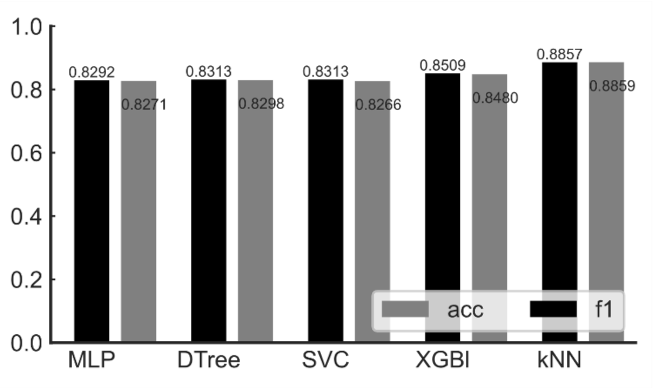


Figure 4. Accuracy and F1 score of all individual learners.

Figure 5 is the confusion matrix of each model on the testing set, where the vertical axis represents the true label of the sample, and the horizontal axis represents the predicted label of the model's predicted output. From the confusion matrix, it can be seen that each model performs well in recognizing class "2". For class "2", the class accuracy of the SVC is highest at 0.9706, and the C-ACC of the MLP is lowest at 0.8725. Each model has different degrees of defects in recognizing class "1". The SVM's ACC on class "1" is only 0.66, and the C-ACC of the kNN is only 0.8511. For kNN, XGBl, and DTree, the recognition errors are relatively evenly distributed among the recognition of classes "0", "1", and "2". The recognition ACC of kNN for each class has reached more than 0.85, and the recognition ACC of class "2" has reached 0.95.

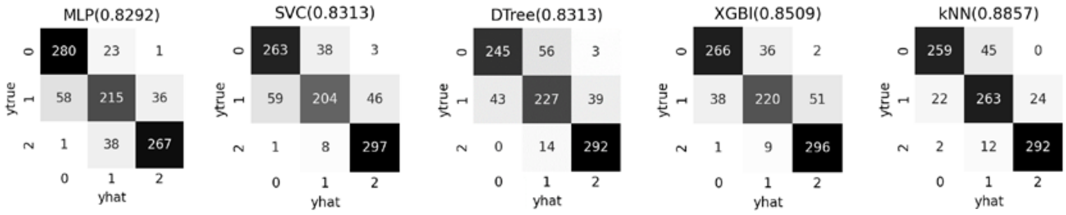


Figure 5. Confusion matrix of all individual learners.

Based on the analysis results of ACC, F1, and testing set confusion matrix, the individual learners are summarized as follows:

(1) kNN and XGBl have high ACC, and the misjudgment results for different classes of soils are relatively evenly distributed, and both have an excellent discriminant effect on soil class "2". So, they two can be used as solid classifiers to complete the L1 integration.

(2) The ACC of the SVM, MLP, and DTree is low, and it is incidental to misjudge the soil of class "1" during the model prediction process, but they perform well in distinguishing the soil of class "2" and class "0" respectively. So, they can be used as weak classifiers to complete the L1 integration.

The five individual learners are trained and integrated as the L1 under the same experimental environment. Individual learner in the L1 predicts the sample and combines all the probabilistic prediction sets into a transition data set S. Then, the 203-dimensional raw data of the training set is output as a 15-dimensional S through the L1. Input S into the L2 alternative model to complete the model fitting analysis.

Table 4 compares the ACC and F1 of the nine stacking models before and after simulated annealing hyperparameter optimization. It can be seen that the ACC of the model after the simulated annealing hyperparameter optimization has been dramatically improved. The Logist-stacking model has the best performance with an ACC of 0.8923, and the DTree-stacking model has the worst performance with an ACC of 0.8651. It can be seen in the table that the inter-class classification ability of each classifier is the same, and the classification effect is similar, and it can complete the multi-classification task of this paper.

Table 4. Accuracy and F1 score before and after optimization of each stacking model parameter.

	Before Optimization		**After Optimization**				
	acct	f1	acct	acc_0	acc_1	acc_2	f1
DTree	0.6047	0.5959	0.8651	0.8520	0.7573	0.9869	0.8628
RF	0.6212	0.5809	0.8694	0.8553	0.7735	0.9804	0.8677
ada	0.5306	0.4467	0.8716	0.8520	0.7832	0.9804	0.8703
XGB	0.6059	0.5248	0.8760	0.8520	0.7961	0.9804	0.8749
LSVM	0.6318	0.5995	0.8760	0.8586	0.8026	0.9673	0.8754
SVC	0.6247	0.5910	0.8792	0.8586	0.7994	0.9804	0.8783
kNN	0.6153	0.5965	0.8825	0.8586	0.8252	0.9641	0.8822
MLP	0.6294	0.6007	0.8879	0.8750	0.8155	0.9739	0.8872
Logist	0.6459	0.6123	0.8912	0.8750	0.8252	0.9739	0.8905

3.2 Applicability verification of model

In order to further verify the ACC of the model in different samples, we introduced a new dataset: the validation set. 2831 sample points in the validation set inputted nine stacking models.

It can be seen from Table 5 that the ACC of the nine stacking models on the validation set is quite different. The LSVM-stacking model has the best performance with an ACC of 0.9488, and the Ada-stacking model has the worst performance with an ACC of 0.6708. In addition, each classifier has significantly improved the recognition effect of class "1", and the recognition effect of class "2" is also slightly improved. Instead, the main classification errors are concentrated in recognizing class "0". The C-ACC of the LSVM-stacking model in both class "1" and class "2" has reached 1.0, and the C-ACC in class "0" has also reached 0.7940, which is much higher than other stacking models, indicating that the LSVM-stacking model has the best applicability in the validation set.

Table 5. Accuracy of each stacking model on validation set.

	acct	**acc_0**	**acc_1**	**acc_2**	**f1**
Ada	0.6708	0.0000	0.8223	0.9641	0.5177
DTree	0.7400	0.0000	0.9701	1.0000	0.5756
XGBl	0.7513	0.0000	1.0000	1.0000	0.5841
Logist	0.7835	0.4091	0.9673	0.8469	0.7483
SVC	0.9022	0.6065	1.0000	1.0000	0.8799
kNN	0.9036	0.6577	0.9963	0.9735	0.8874
RF	0.9050	0.6335	0.9897	1.0000	0.8842
MLP	0.9131	0.6506	1.0000	1.0000	0.8930
LSVM	0.9488	0.7940	1.0000	1.0000	0.9401

Finally, LSVM is determined to be an L2 function.

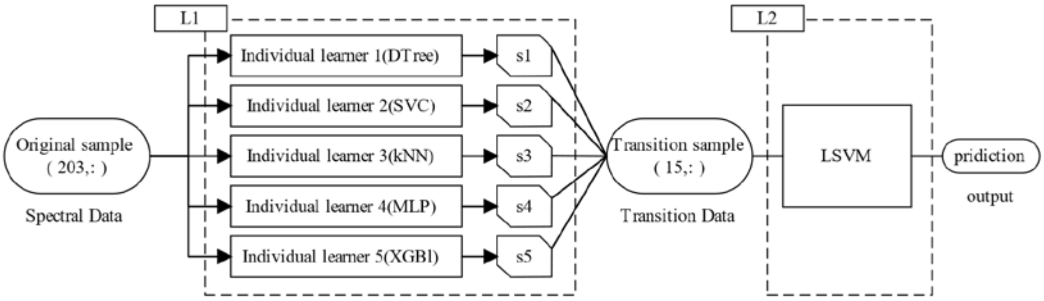

L1
Individual learner 1(DTree)
Individual learner 2(SVC)
Individual learner 3(kNN)
Individual learner 4(MLP)
Individual learner 5(XGBI)
s1
s2
s3
s4
s5
Original sample (203,:)
Spectral Data
Transition sample (15,:)
Transition Data
L2
LSVM
pridiction
output

Figure 6. The final process structure of the algorithm.

4. Discussion

The representative XGBoost ensemble learning algorithm is an excellent machine learning algorithm with better stability than a single individual learner, but not all ensemble learning algorithms can complete the classification task well. The stacking model proposed in this paper is not entirely based on the tree model, and its final performance is not necessarily better than the traditional classifier model. However, using more types of individual learners in the L1 can better consider the diversity of the data set, improving the model's applicability ability. Greater adaptability is why a brand-new validation set is introduced to validate the model. If we want to choose a simple classification model, kNN can do the job well. When dealing with hyperspectral data, the data dimension is high, and there is much redundant information. It is necessary to filter the data information through L1. The preliminary judgment result of L1 is given to L2 for analysis and judgment, which also can reduce the workload of L2 functions.

5. Conclusions

In this paper, an ensemble learning model based on black soil hyperspectral data is designed to complete the classification of SOM content classification quickly. Five individual learners are selected as L1 to improve the applicability ability of the model, and the simulated annealing algorithm is used to complete the hyperparameter optimization of the LSVM-stacking model after 500 iterations. The regularization parameter of the best model is C=0.34. The ACC on the testing set before the parameter adjustment is 0.6318, and the ACC on the testing set after the parameter adjustment is 0.8760. In the independent validation set data, the ACC of the LSVM-stacking model is 0.9488 higher than other stacking models, which can improve the C-ACC of class "1" to 1.0. The classification ability of this stacking model is more balanced than before the parameter adjustment, and the F1 has also improved accordingly.

Author Contributions: Conceptualization, Z.Z. and J.F.; Methodology, Z.Z.; Software, Z.Z. and Q.Z.; Validation, Z.Z. and Z.L.; formal analysis, Z.L.; data curation, Z.Z. and H.X.; Writing – original draft, Z.Z.; Writing – review & editing, Z.Z., K.T. and J.F.

Funding: This research was funded by the Natural Science Foundation of Heilongjiang Province, grant number LH2020C003.

Institutional Review Board Statement: Not applicable. For studies not involving humans or animals.

Data Availability Statement: The dataset and fitted models can be downloaded at the GitHub at: https://github.com/rusuanjunccdong/Three-grade-Black-Soil.

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

References

1. Liao, Q.; Gu, X.; Li, C.; Chen, L.; Huang, J.; Du, S.; Fu, Y.; Wang, J. Estimation of fluvo-aquic soil organic matter content from hyperspectral reflectance based on continuous wavelet transformation. *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)* **2012**, 132-139+298, doi:10.3969/j.issn.1002-6819.2012.23.018.
2. No.4[2022] Document of State Council Notice of the State Council on Launching the Third National Soil Survey.
3. Hu, Y.; Jiang, H.; Zhuo, H.; Xu, L.; Ju, H.; Wang, Y. Study on Detection Method of Maturity of Camellia Oleifera Based on Hyperspectral Imaging. *Food Science* 1–13, doi:10.7506/spkx1002-6630-20210619-229.
4. Guo, Z.; Jin, C.; Liu, P.; Tang, X.; Zhao, N. Research Progress of Spectral Analysis and Spectral Imaging Technology in Soybean Quality Detection. Soybean Science 2022, 41, 99–106, doi:10.11861/j.issn.1000-9841.2022.01.0099.
5. Fan, L.; He, L.; Tan, C.; Tian, Y.; Zhang, C.; Wu, C.; Huang, Y. Rapid Detection of Adulteration of Fritillariae Cirrhosae Bulbus Based on Portable Near Infrared Spectroscopy. Chinese Journal of Experimental Traditional Medical Formulae 1–13, doi:10.13422/j.cnki.syfjx.20211757.
6. Liu, Y.; Tan, K.; Chen, Y.; Wang, Z.; Xie, H.; Wang, L. Variety Recognition of Soybeans Using Segmented Principal Component Analysis and Hyperspectral Technology. Soybean Science 2016, 35, 672–678, doi:10.11861 /j.issn.1000-9841.2016.04.0672.
7. Chai, Y.; Bi, W.; Tan, K.; Zhang, C.; Liu, C. Nondestructive Identification of Soybean Seed Varieties Based on Hyperspectral Image Technology. Journal of Northeast Agricultural University 2016, 47, 86–93.
8. Jia, A.; Dong, T.; Zhang, Y.; Zhu, B.; Sun, Y.; Wu, Y.; Shi, Y.; Ma, Y.; Guo, Y. Recognition of Field-Grown Tobacco Plant Type Characteristics Based on Three-Dimensional Point Cloud and Ensemble Learning. Journal of Zhejiang University (Agric. & Life Sci.) 1–10, doi:10.3785/j.issn.1008-9209.2021.05.173.
9. Yi, X.; Zhang, L.; Lv, X.; Zhang, Z.; Tian, M.; Yin, C.; Ma, Y.; Fan, X. Estimation of Cotton Above-Ground Biomass Based on Unmanned Aerial Vehicle Hyperspectral and Successive Projections Algorithm. Cotton Science 2021, 33, 224–234, doi:10.11963/1002-7807.yxzlf.20210428.
10. Wang, L.; Wang, L. Variety Identification Model for Maize Seeds Using Hyperspectral Pixel-Level Information Combined with Convolutional Neural Network. National Remote Sensing Bulletin 2021, 25, 2234–2244.
11. Bai, Q.; Hou, Y.; Yang, P.; Li, W.; Lin, X.; Su, J.; Xu, J.; Liu, X. Identification Method of the Producetion Site of Gastrodia Elata Blume Based on Near Infrared Spectroscopy. Journal of West China Forestry Science 2021, 50, 124–130.
12. Zhu, S.; Chao, M.; Zhang, J.; Xu, X.; Song, P.; Zhang, J.; Huang, Z. Identification of Soybean Seed Varieties Based on Hyperspectral Imaging Technology. Sensors (Basel, Switzerland) 2019, 19, 5225.
13. Yin, K.; Liu, J.; Zhang, D.; Zhang, A. Rapiddetectionofproteincontentinricebasedon Nearinfraredspectroscopy. FOOD&MACHINERY 2021, 37, 82-88+175.
14. Gao, H.; Wang, G.; Zhao, J.; Wang, Z. Model Optimization for Determination of Common Nutritional Components in Wheat Flour Using Near Infrared Spectroscopy. Food and Nutrition in China 2021, 27, 30–34.
15. Yang, M.; Yu, J.; Huang, Y. Nondestructive Detection for Apple Quality Using Near-Infrared (NIR) Spectroscopy: A Ｒeview. FORESTRY MACHINERY ＆ WOODWORKING EQUIPMENT 2021, 49, 4–8.
16. Liu, H.; Bao, Y.; Meng, X.; Cui, Y.; Zhang, A.; Liu, Y.; Wang, D. Inversion of Soil Organic Matter Based on GF-5 Images under Different Noise Reduction Methods. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE) 2020, 36, 90–98, doi:10.11975/j.issn.1002-6819.2020.12.011.
17. Zhu, Y.; Wang, D.; Zhang, H.; Shi, P. Soil Organic Carbon Content Retrieved by UAV-Borne High Resolution Spectrometer. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE) 2021, 37, 66–72, doi:10.11975/j.issn.1002-6819.2021.06.009.
18. Ye, H.; Xiong, H.; Zhang, F.; Wang, N.; Ma, L. CWT-Based Estimation of Soil Organic Matter Content in Arid Area Under Different Human Disturbance Degrees. Laser& Optoelectronics Progress 2019, 56, 115–124.

19. Wang, D.; Qin, K.; Li, L.; Zhao, Y.; Chen, W.; Gan, Y. Retrieval of Organic Matter Content in Black Soil Based on Airborne Hyperspectral Remote Sensing Data: Taking Jiansanjiang District in Heilongjiang Province as an Example. earth science 2018, 43, 2184–2194.
20. Tao, P.; Wang, J.; Li, Z.; Zhou, P.; Yang, J.; Gao, F. RESEARCH OF SOIL NUTRIENT CONTENT INVERSION MODEL BASED ON HYPERSPECTRAL DATA. GEOLOGY AND RESOURCES 2020, 29, 68-75+84.
21. Tang, M.; Yang, Q.; Tang, H. Estimation of Soil Organic Carbon Content Using Ypersp Ectral Data in Peak-Ciuster Depressions, Northeastern Guangxi. Carsologica Sinica 1–9.
22. Reis, A.S.; Rodrigues, M.; Alemparte Abrantes dos Santos, G.L.; Mayara de Oliveira, K.; Furlanetto, R.H.; Teixeira Crusiol, L.G.; Cezar, E.; Nanni, M.R. Detection of Soil Organic Matter Using Hyperspectral Imaging Sensor Combined with Multivariate Regression Modeling Procedures. Remote Sensing Applications: Society and Environment 2021, 22, 100492, doi:10.1016/j.rsase.2021.100492.
23. Zhou, Z. Machine learning; Tsinghua university press, 2016; ISBN ISBN 978-7-302-206853-6.
24. Ceamanos, X.; Waske, B.; Benediktsson, J.A.; Chanussot, J.; Fauvel, M.; Sveinsson, J.R. A Classifier Ensemble Based on Fusion of Support Vector Machines for Classifying Hyperspectral Data. International Journal of Image and Data Fusion 2010, 1, 293–307, doi:10.1080/19479832.2010.485935.
25. Huang, X.; Zhang, L. An SVM Ensemble Approach Combining Spectral, Structural, and Semantic Features for the Classification of High-Resolution Remotely Sensed Imagery. IEEE Trans. Geosci. Remote Sensing 2013, 51, 257–272, doi:10.1109/TGRS.2012.2202912.
26. Jonathan Cheung-Wai Chan; Chengquan Huang; DeFries, R. Enhanced Algorithm Performance for Land Cover Classification from Remotely Sensed Data Using Bagging and Boosting. IEEE Trans. Geosci. Remote Sensing 2001, 39, 693–695, doi:10.1109/36.911126.
27. Pal, M.; Mather, P.M. An Assessment of the Effectiveness of Decision Tree Methods for Land Cover Classification. Remote Sensing of Environment 2003, 86, 554–565, doi:10.1016/S0034-4257(03)00132-9.
28. Chan, J.C.-W.; Paelinckx, D. Evaluation of Random Forest and Adaboost Tree-Based Ensemble Classification and Spectral Band Selection for Ecotope Mapping Using Airborne Hyperspectral Imagery. Remote Sensing of Environment 2008, 112, 2999–3011, doi:10.1016/j.rse.2008.02.011.
29. Ham, J.; Yangchi Chen; Crawford, M.M.; Ghosh, J. Investigation of the Random Forest Framework for Classification of Hyperspectral Data. IEEE Trans. Geosci. Remote Sensing 2005, 43, 492–501, doi:10.1109/TGRS.2004.842481.
30. Guo, L.; Sun, X.; Fu, P.; Shi, T.; Dang, L.; Chen, Y.; M, L.; Zhang, G.; Zhang, Y.; Jiang, Q.; et al. Mapping Soil Organic Carbon Stock by Hyperspectral and Time-Series Multispectral Remote Sensing Images in Low-Relief Agricultural Areas. Geoderma 2021, Volume 398,.
31. Guo, D.; Zhai, J.; Xie, X.; Zhu, Y. Heterogeneous Ensemble Spectral Classifiers for Hyperspectral Images. Procedia Computer Science 2021, 187, 229–234, doi:10.1016/j.procs.2021.04.115.
32. DZ_T 0279.27-2016 区域地球化学样品分析方法;
33. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics 1953, 21, 1087–1092, doi:10.1063/1.1699114.
34. Vecchi, M.P.; Kirkpatrick, S. Global Wiring by Simulated Annealing. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 1983, 2, 215–222, doi:10.1109/TCAD.1983.1270039.
35. Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P. Optimization by Simulated Annealing. Science 1983, 220, 671–680, doi:10.1126/science.220.4598.671.
36. Černý, V. Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm. J Optim Theory Appl 1985, 45, 41–51, doi:10.1007/BF00940812.

37. Yao, X.; Chen, G. Simulated Annealing Algorithm and Its Applications. Journal Of Computer Research and Development 1990, 1–6.