

Article

Estimating Soil Cd Contamination in Wheat Farmland Using Hyperspectral Data and Interpretable Stacking Ensemble Learning

Liang Zhong ^{1,†}, Meng Ding ^{2,†}, Shengjie Yang ¹, Xindan Xu ³, Jianlong Li ^{1,*} and Zhengguo Sun ^{4,*} 

¹ Department of Ecology, School of Life Sciences, Nanjing University, Nanjing 210023, China; zhongliang@smail.nju.edu.cn (L.Z.)

² Jiangsu Environmental Protection Group Suzhou Co., Ltd., Suzhou 215009, China

³ College of Life Sciences, Shangxi University, Taiyuan 030031, China

⁴ College of Agro-Grassland Science, Nanjing Agricultural University, Nanjing 210095, China

* Correspondence: jlli2008@nju.edu.cn (J.L.); sunzg@njau.edu.cn (Z.S.)

† These authors contributed equally to this work.

Abstract

Soil heavy metal pollution threatens agricultural safety and human health, with Cd exceeding standards being the most common problem in contaminated farmland. The development of hyperspectral remote sensing technology has provided a novel methodology of quickly and non-destructively monitoring heavy metal contamination in soil. This study aims to explore the potential of an interpretable Stacking ensemble learning model for the estimation of soil Cd contamination in farmland hyperspectral data. We assume that this method can improve the modeling accuracy. We chose Zhangjiagang City, Jiangsu Province, China, as the study area. We gathered soil samples from wheat fields and analyzed soil spectral data and Cd level in the lab. First, we pre-processed the spectra utilizing fractional-order derivative (FOD) and standard normal variate (SNV) transforms to highlight the spectral features. Second, we applied the competitive adaptive reweighted sampling (CARS) feature selection algorithm to identify the significant wavelengths correlated with soil Cd content. Then, we constructed and compared the estimation accuracy of multiple machine learning models and a Stacking ensemble learning method and utilized the Optuna method for hyperparameter optimization. Ultimately, the SHAP method was used to shed light on the model's decision-making process. The results show that (1) FOD can further highlight the spectral features, thereby strengthening the correlation between soil Cd content and wavelength; (2) the CARS algorithm extracted 3.4–6.8% of the feature wavelengths from the full spectrum, and most of them were the wavelengths with high correlation with soil Cd; (3) the optimal estimation precision was achieved using the FOD1.5-SNV spectral pre-processing combined with the Stacking model ($R^2 = 0.77$, RMSE = 0.05 mg/kg, RPD = 2.07), and the model effectively quantitatively estimated soil Cd contamination; and (4) SHAP further revealed the contribution of each base model and characteristic wavelengths in the Stacking modeling process. This research confirms the advantages of the interpretable Stacking model in hyperspectral estimation of Cd contamination in farmland wheat soil. Furthermore, it offers a foundational reference for the future implementation of quantitative and non-destructive regional monitoring of heavy metal contamination in farmland soil.

Keywords: soil heavy metal; Cd contamination; hyperspectral; stacking ensemble learning; SHAP; fractional-order derivative



Academic Editor: Pablo Martín-Ramos

Received: 31 May 2025

Revised: 25 June 2025

Accepted: 26 June 2025

Published: 27 June 2025

Citation: Zhong, L.; Ding, M.; Yang, S.; Xu, X.; Li, J.; Sun, Z. Estimating Soil Cd Contamination in Wheat Farmland Using Hyperspectral Data and Interpretable Stacking Ensemble Learning. *Agronomy* **2025**, *15*, 1574. <https://doi.org/10.3390/agronomy15071574>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Soil contamination poses a significant threat to global agricultural security and human health [1,2]. Recent studies have shown that about 14–17% of the agricultural land in the world has at least one toxic metal in excess, with a significant excess of Cd [3]. Cd is a highly toxic heavy metal that exhibits high mobility, bioaccumulation, and long-term residue. It is easily assimilated by plants and subsequently accumulates within the food chain, thereby presenting a significant risk to both ecosystems and human health [4,5]. Traditional soil Cd detection relies on laboratory chemical analysis, which is highly accurate. However, there are defects such as complicated sample pretreatment, a long cycle time, high cost, and difficulty in obtaining spatial distribution information over a large area [6]. These issues prevent it from meeting the needs of large-scale dynamic monitoring. Recently, hyperspectral remote sensing technology has emerged as an innovative method for the rapid and non-destructive assessment of heavy metal concentrations in soil [7,8]. This advancement can be attributed primarily to its numerous spectral bands and high spectral resolution. The hyperspectral technique can quickly monitor soil Cd contamination by capturing detailed spectral characteristics of soil in the visible and near-infrared ranges. This method indirectly indicates the Cd levels and its interactions with iron oxides, organic matter, and clay minerals [9]. Nevertheless, hyperspectral data has extremely high dimensionality (hundreds to thousands of bands), a high band correlation, and interference from ambient noise [10]. This results in a serious challenge to extract effective information from hyperspectral data and construct robust estimation models. Competitive adaptive reweighted sampling (CARS) is a feature selection method based on variable competitive screening. Research by Liu et al. [11] and Li et al. [12] shows that CARS is an efficient method for extracting sensitive spectral information and reducing redundant bands.

Since the inherent noise (e.g., scattering effect, baseline drift) of hyperspectral data can mask the target signal, the effective information needs to be enhanced by pre-processing [13]. The derivative transform is a widely used spectral pre-processing method. For example, researchers often utilize first-order or second-order derivatives to mitigate baseline drift, resolve overlapping peaks, and enhance minor changes in spectral features [14]. This approach is undertaken to enhance the precision of estimations regarding heavy metal content in soil. The fractional-order derivative (FOD) represents an extension of the integer-order derivative, thereby providing a more detailed treatment of the spectral derivative. Finer quantization on the order means that curvature and tilt variations in each band can be captured more sensitively, highlighting spectral features [15,16]. For instance, Chen et al. [17] and Cao et al. [18] studied optimization of the order of the FOD by optimizing the order of the fractional-order derivatives in fine steps (e.g., 0.1 or 0.25) between the integer-order derivatives. They found the most sensitive spectral information for soil heavy metal elements, which subsequently enhanced the predictive accuracy and reliability of the model. Furthermore, various spectral pre-processing techniques, including derivative transformation and standard normal variate (SNV) transformation, have been integrated by some researchers to maximize the benefits of different spectral transformation approaches. This combination serves to highlight spectral characteristics and enhance the accuracy of modeling [19,20].

The advancement of computer technology has facilitated the extensive application of diverse machine learning algorithms in the analysis of hyperspectral data [21]. Qi et al. [22,23] successfully employed machine learning models, including random forest (RF), XGBoost, LightGBM (LGBM), and AdaBoost, to rapidly identify heavy metal Ni and Cu contamination in the European LUCAS spectral library. Zhou et al. [24] modeled the hyperspectral estimation of various soil heavy metal elements (Cr, Pb, Ni, Cu, Zn, and Mn) in the Sanjiangyuan region of China by RF, support vector regression (SVR), and partial

least squares regression (PLSR) methods. However, most of the existing studies rely on a single model, whose performance is easily constrained by data distribution and parameter settings, making it difficult to break through the accuracy bottleneck. For example, SVR is highly sensitive to kernel functions and penalty coefficients, while RF is prone to overfitting when features are redundant [25,26]. In machine learning, ensemble learning is defined as a method that utilizes the prediction results of several base models to enhance overall accuracy. The key idea is to reduce variance and bias through diversity among models (e.g., algorithmic differences, data subsets, or feature subsets) to improve prediction accuracy and robustness [27]. The Stacking ensemble learning model integrates the prediction results of multiple base models by training a meta-model. This approach comprehensively utilizes the advantages inherent in each base model while concurrently circumventing the limitations associated with a single model [28,29]. However, complex ensemble learning models are usually perceived as “black boxes”, which makes it difficult to understand how they make decisions. This lack of interpretability constrains their applicability in real-world scenarios [30].

In the domains of environmental monitoring and agriculture, the interpretability of models is of paramount importance. Models that are interpretable facilitate a deeper grasp of the decision-making mechanisms underlying predictions, thereby fostering greater trust among researchers and decision-makers in the reliability of these predictions [31]. The SHapley Additive exPlanations (SHAP) methodology, which is grounded in game theory, serves as a model interpretation technique that facilitates the allocation of a value to each feature [32,33]. The visualization of SHAP values enables researchers to gain an intuitive comprehension of the features that exert a substantial influence on the predictions made by the model. This practice enhances the transparency and credibility of the model's outcomes.

This study aims to examine the validity of interpretable Stacking ensemble learning models for hyperspectral estimation of Cd contamination in farmland soil. Furthermore, the value of interpretable ensemble learning in the hyperspectral estimation of soil contamination will be clarified. First, we pre-processed the spectra using FOD and SNV transforms to highlight spectral features. Second, we applied the CARS feature selection algorithm to identify the significant wavelengths associated with soil Cd content. Then, we constructed and compared the estimation performance of multiple machine learning models and the Stacking model. Among them, we used simple, efficient, and widely used PLSR as the meta-model. Ultimately, the SHAP method was used to shed light on the model's decision-making process.

2. Materials and Methods

2.1. Study Area

Zhangjiagang City (31°43′–32°02′ N, 120°21′–120°52′ E) is situated in the southeastern part of Jiangsu Province, China, as well as on the southern bank of the lower Yangtze River (Figure 1). The city is an early open port industrial city in the intersection area of the coastal and Yangtze River economic zones and covers a total area of about 989 km². The area is situated within the southern humid climate zone of the northern subtropical region, which is characterized by a temperate climate, the presence of four distinct seasons, and a significant amount of precipitation. The average annual temperature, precipitation, and sunshine hours are 16 °C, 1078 mm, and 1887 h, respectively. The terrain of the whole region is flat, with the elevation located from −8 m to 135 m, which is suitable for large-scale farming. As a result, agriculture in Zhangjiagang City is well developed, with a cultivated area of 313 km², which is mainly cultivated in two seasons a year, with wheat and rice in rotation. During the development of industrialization, some farmlands are vulnerable

to the influence of industrial and domestic wastewater, atmospheric dust fall, etc., which adversely affect the soil environment of farmlands.

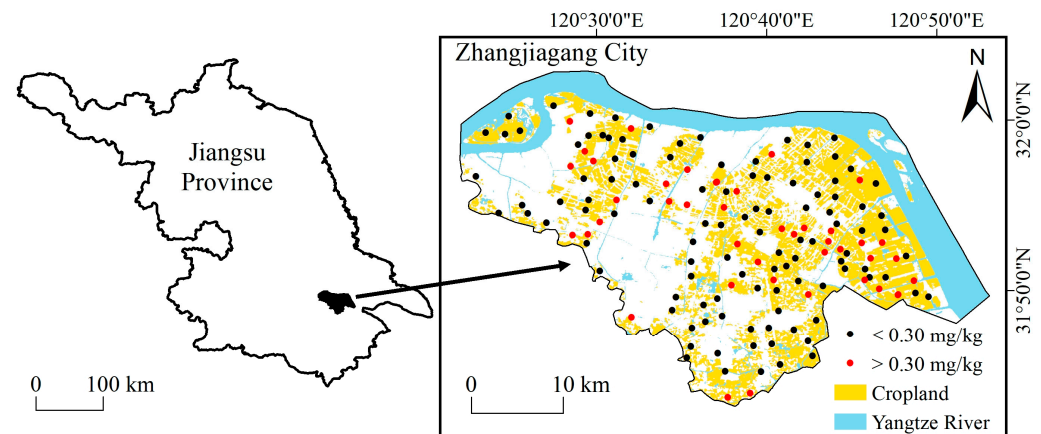


Figure 1. Location of study area and sampling points.

2.2. Data Collection and Determination

During 16–21 May 2024, we sampled 152 soil samples from the 0–20 cm surface layer of farmland wheat in the study area (Figure 1). At the time of sampling, we collected soil samples (approximately 1 kg) using the five-point mixing method at each sample point and brought the samples back to the laboratory in sealed bags [24]. At the same time, we recorded the number, location, and surrounding environment information of each sampling point.

In the lab, the soil samples were subjected to a series of processing steps. Firstly, they were air-dried, then ground, and subsequently sieved. The processed samples were then split into two distinct parts. A portion of the samples was utilized for the purpose of determining the soil Cd content by ICP-MS [34]. An additional portion of the samples was utilized to determine soil spectral data through the utilization of an ASD FieldSpec3 ground spectrometer (Analytical Spectral Devices Inc., Boulder, CO, USA). The instrument operates within a wavelength range of 350 to 2500 nm, a resampling interval of 1 nm, and 2151 bands of data per spectral output. The scans were repeated five times for each sample, and the average value was calculated to represent the soil spectral result for that sample. During the measurement process, calibration was carried out every 10 min using a standard white plate.

2.3. Research Methods

2.3.1. Workflow

The workflow of this study is illustrated in Figure 2. (1) Data collection and determination: soil samples were obtained from agricultural fields, and subsequent laboratory analyses were conducted to determine the content of soil Cd as well as to acquire soil spectral data. (2) Spectral pre-processing: the spectra were pre-processed using FOD and SNV. (3) Feature selection: CARS was utilized to select the feature wavelengths of soil Cd. (4) Modeling: a Stacking ensemble learning method consisting of multiple machine learning models was constructed, and the modeling performance was evaluated. (5) Model interpretation: the contributions of the base models and feature wavelengths were calculated using the SHAP method.

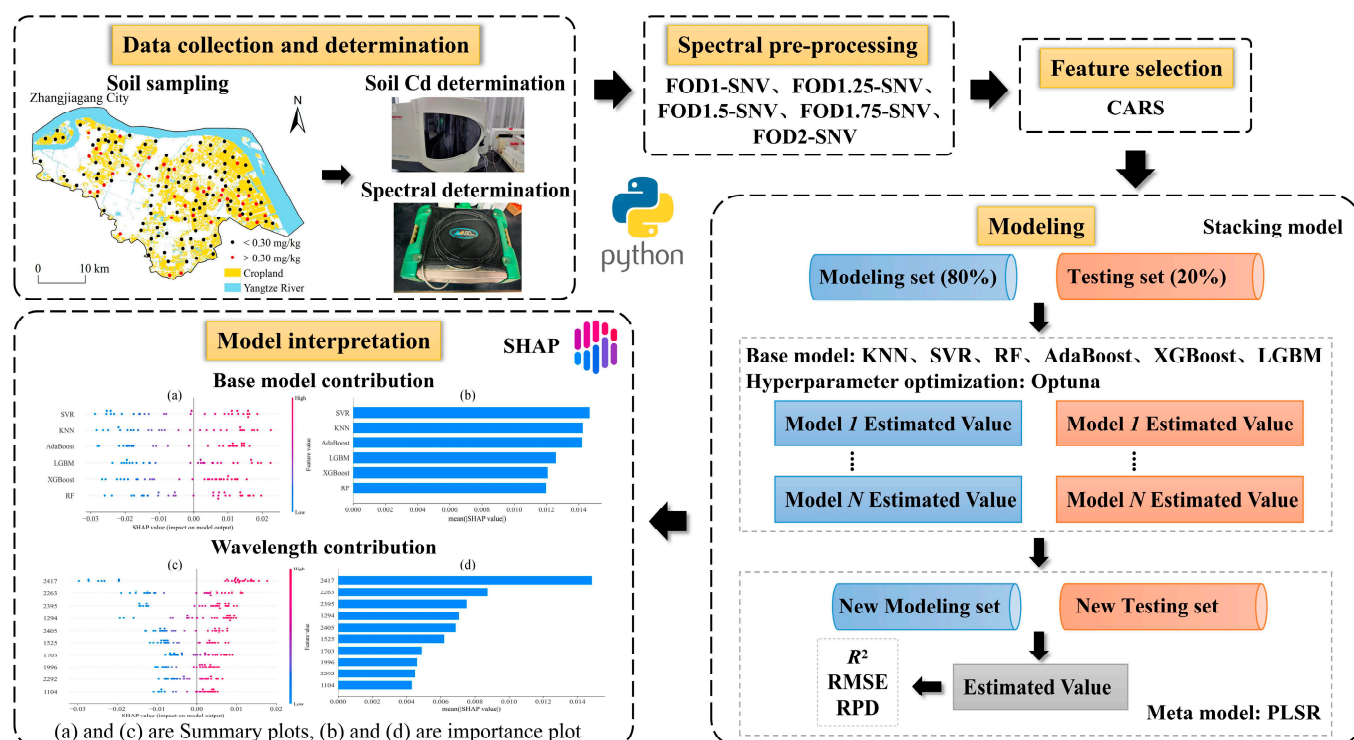


Figure 2. Flowchart for estimating Cd contamination in farmland soil.

2.3.2. Spectral Pre-Processing

Initially, the raw spectral data underwent a Savitzky–Golay smoothing process and then further FOD transformed using the Grünwald–Letnikov algorithm [35]. This resulted in 1-order (FOD1), 1.25-order (FOD1.25), 1.5-order (FOD1.5), 1.75-order (FOD1.75), and 2-order (FOD2) transformed spectra. Then, for each FOD spectrum, the SNV transform process was performed. Finally, the spectral bands within the ranges of 350–400 nm and 2451–2500 nm, characterized by low signal-to-noise ratios, were excluded from the analysis. Conversely, the bands spanning from 401 to 2450 nm were preserved for the purposes of spectral analysis and modeling at each sampling location.

2.3.3. Feature Selection

CARS is a feature selection technique that employs a competitive screening process for variables. It is extensively utilized for identifying critical wavelengths in the modeling of high-dimensional spectral data [36]. The core idea is to gradually screen out the subset of features that contribute most to the regression or classification task through an iterative process. At the same time, redundant and irrelevant features are removed to enhance both the performance and computational efficiency of the model. The CARS algorithm is operated in the following steps. First, the algorithm initializes the feature weights and samples the original dataset. In each iteration, the significance of each feature in relation to the target variable is assessed through the computation of the weighted correlation coefficient associated with that feature. Then, the features are ranked according to their importance, and the feature weights are adjusted by an adaptive reweighting strategy. This process is repeated until a preset stopping condition is met, such as when the number of features attains a specified threshold or when the number of iterations reaches a defined upper limit. Eventually, the algorithm outputs a filtered subset of features for subsequent modeling and analysis [37]. In the present study, the number of iterations was established at 50.

2.3.4. Spectral Modeling

Stacking ensemble learning represents a method approach within ensemble techniques aimed at enhancing both the performance and the robustness of models through the integration of predictive outcomes derived from various foundational models [38]. The core idea is divided into two layers. The first layer comprises several heterogeneous base models that have been trained in parallel to produce initial predictions on the original dataset. The second layer retrains the primary predictions by a meta-model to learn the optimal combinatorial relationship among the base models. The base models selected in this study include K-nearest neighbors (KNN), SVR, RF, AdaBoost, XGBoost, and LGBM. The meta-model is the PLSR algorithm, which is capable of efficiently handling high-dimensional collinear data. In addition, it extracts key information from the output of the base model by projective dimensionality reduction, which enhances the generalization capability [39].

KNN is a simple and classical nonparametric supervised learning algorithm, whose core idea is to make predictions based on the similarity of samples. For unknown samples, the K-nearest neighbor samples are selected by calculating the distance to all samples within the training dataset. The prediction of the unknown samples is then determined by voting or weighted averaging based on the category (classification task) or attribute values (regression task) of these neighbors [40].

SVR is a supervised learning algorithm grounded in statistical learning theory. It facilitates the prediction of the target variable by identifying a hyperplane that maximizes the margin, thereby transforming input features into a high-dimensional space. SVR has good generalization ability and is able to effectively handle nonlinear relationships as well as high-dimensional datasets [41].

RF is an ensemble learning algorithm that utilizes decision tree methodologies. RF improves the robustness of models by generating multiple decision trees and employing ensemble learning techniques to aggregate their predictions. The algorithm is designed to construct each tree by randomly selecting a subset of samples and a subset of features. This process introduces randomness, thereby reducing the variance of the model and enhancing generalization. Consequently, this approach effectively avoids overfitting [42].

AdaBoost is an iterative ensemble learning algorithm. A strong learner is constructed by combining multiple weak learners. The main idea is to change the weights of the samples gradually through an iterative approach, which makes the subsequent weak learners to focus more intently on the samples that were incorrectly predicted by the previous model [43].

XGBoost is an ensemble learning algorithm that utilizes the principles of gradient boosting. By using second-order Taylor expansion to approximate the loss function, XGBoost also adds a regularization term to manage the model's complexity and prevent overfitting. In addition, XGBoost supports parallel computing and distributed training, which can effectively improve the training speed [44].

LGBM is a machine learning algorithm that has been developed by Microsoft and is based on the gradient boosting framework. LGBM significantly improves the speed of model training and memory usage of the model by employing a histogram-based decision tree algorithm and a series of optimization techniques, such as one-sided gradient sampling and mutually exclusive feature bundling [45].

2.3.5. Hyperparameter Optimization

Optuna is an effective framework for hyperparameter optimization that combines a dynamic pruning mechanism with an intelligent sampling strategy. In the sampling phase, Optuna adopts the tree-structured parzen estimator algorithm to dynamically predict the

hyperparameter combinations that are more likely to produce optimal results by modeling the results of historical trials [46]. This method significantly improves the search efficiency compared to traditional techniques, including random search and grid search. At the same time, Optuna introduces an early-stop mechanism, which terminates the experiment in time when the model performance during the experiment is significantly worse than the historical optimum, avoiding ineffective computational resource consumption. The advantages of Optuna are as follows: first, it is highly automated and scalable, supports various machine learning frameworks, and is able to adapt to the structure of complex models; second, through parallelization and distributed computing, the optimization time is significantly reduced; third, it is rich in visualization tools, which makes it easy for researchers to intuitively analyze the relationship between hyperparameters and the objective function [28].

In the present research, the Optuna methodology was utilized for the optimization of hyperparameters across several machine learning models (Table 1). The objective function was a 5-fold cross-validated root mean square error (RMSE) with a termination condition of 100 iterations of optimization. All modeling processes were run on a laptop with a Windows 11 operating system equipped with an i7-12700H CPU with 16 GB of memory and an NVIDIA GeForce RTX 3060 GPU with 6 GB of memory, and the modeling was completed using the Python 3.9 (<https://www.python.org/> (accessed on 8 April 2025)) language environment. The modeling times for KNN, SVR, RF, AdaBoost, XGBoost, LGBM, and Stacking were approximately 5 s, 1 s, 100 s, 90 s, 80 s, 10 s, and 290 s, respectively.

Table 1. Model hyperparameter optimization range settings.

Model	Parameters	Range	Step Size
KNN	n_neighbors	(1, 10)	1
	leaf_size	[1, 5, 10, 20, 30, 40, 50]	
SVM	p	(1, 10)	1
	C	[1, 5, 10, 50, 100, 500, 1000]	
RF	gamma	['scale', 'auto']	
	n_estimators	[50, 100, 150, 200]	
	max_depth	(2, 10)	1
	min_samples_split	(2, 10)	1
AdaBoost	min_samples_leaf	(1, 10)	1
	n_estimators	[50, 100, 150, 200]	
	learning_rate	[0.001, 0.01, 0.1, 1]	
XGBoost	n_estimators	[50, 100, 150, 200]	
	max_depth	(2, 10)	1
	num_leaves	(10, 50)	10
LGBM	learning_rate	[0.001, 0.01, 0.1, 1]	
	n_estimators	[50, 100, 150, 200]	
	max_depth	(2, 10)	1
	num_leaves	(10, 50)	10
	learning_rate	[0.001, 0.01, 0.1, 1]	
	reg_lambda	[0.001, 0.01, 0.1, 1]	

2.3.6. Model Evaluation

We chose the coefficient of determination (R^2), RMSE, and relative percent deviation (RPD) for modeling accuracy evaluation. Among them, a value of R^2 that approaches 1 indicates superior model fit and stability, while a lower root mean square error (RMSE) signifies enhanced model accuracy. When the RPD value exceeds 2.00, this signifies that the model is capable of being quantitatively estimated. When the RPD value is between 1.50 and 2.00, this suggests that the model can be estimated with a degree of approximation.

When the RPD value falls below 1.50, this indicates that the model exhibits inadequate estimation capability [47].

2.3.7. Model Interpretation

SHAP is a model interpretation method that is based on the concept of the Shapley value, as it appears in the field of game theory [48]. The purpose of this method is to quantify the contribution of each feature to the prediction of the model. The core idea is to consider the prediction result as the “game gain” of all features working together and to assign importance through the calculation of the marginal contribution of features in different combinations. SHAP values ensure fairness of interpretation, satisfying additivity (the total sum of SHAP values of all features is equivalent to the deviation of the predicted value from the baseline value) and consistency (the ordering of feature importance is consistent with the model behavior). Feature importance ranking is consistent with model behavior. Its advantages include the following: (1) the method can analyze the prediction basis of individual samples as well as assess the overall feature importance, realizing the unification of global and local interpretations; (2) the method is model-independent and applicable to any machine learning model; (3) the method explicitly shows the positive or negative impact of features on prediction, enhancing interpretability; (4) the method is based on game theory, providing stable and reliable contribution assignments [49].

3. Results

3.1. Soil Sample Statistics

We randomly divided the total dataset into an 80% modeling set (122 samples) and a 20% testing set (30 samples) (Table 2). During the training process, the modeling set underwent 5-fold cross-validation to obtain the optimal hyperparameters. In the total dataset, the range of soil Cd content was 0.004–0.438 mg/kg, with a mean value of 0.221 mg/kg, a standard deviation of 0.109 mg/kg, and a coefficient of variation of 0.492. According to the risk control standards of soil contamination in agricultural land, a total of 43 sampling points within the dataset surpassed the risk screening threshold of 0.3 mg/kg for farmland, accounting for 28%. These over-standard sample points are mainly distributed in the eastern and central regions of the study area (Figure 1). This indicates that there is a certain degree of pollution risk in wheat soil in the study area. The statistical characteristics of the modeling and testing sets were basically consistent with the total dataset, with mean values of 0.221 mg/kg and 0.219 mg/kg and standard deviations of 0.110 mg/kg and 0.106 mg/kg, respectively. This indicates that the divided subset can better represent the distribution characteristics of the original dataset, thereby fulfilling the requirements for model construction and evaluation.

Table 2. Statistical characteristics of soil Cd content in different sample types.

Dataset	Sampling Points	Minimum (mg/kg)	Maximum (mg/kg)	Mean (mg/kg)	Standard Deviation (mg/kg)	Coefficient of Variation
Total set	152	0.004	0.438	0.221	0.109	0.492
Modeling set	122	0.004	0.438	0.221	0.110	0.497
Testing set	30	0.007	0.381	0.219	0.106	0.482

3.2. Spectral Transformation Features

The results obtained from the combination of different FOD transforms with SNV pre-processing are illustrated in Figure 3. The figure shows that the detailed features of the spectral curves are gradually enhanced with the increase in the fractional-order derivative order. Among them, the low-order derivatives (Figure 3a) retain the overall trend of the

original spectra, while the high-order derivatives (Figure 3e) significantly highlight the local absorption peaks and minor fluctuations. This helps to extract finer spectral information. The SNV transform further eliminates baseline drift and scattering interference, making the spectral data more consistent with a normal distribution. The combination of different orders reflects the flexibility of FOD in balancing noise suppression and feature extraction, providing diverse input features for subsequent modeling.

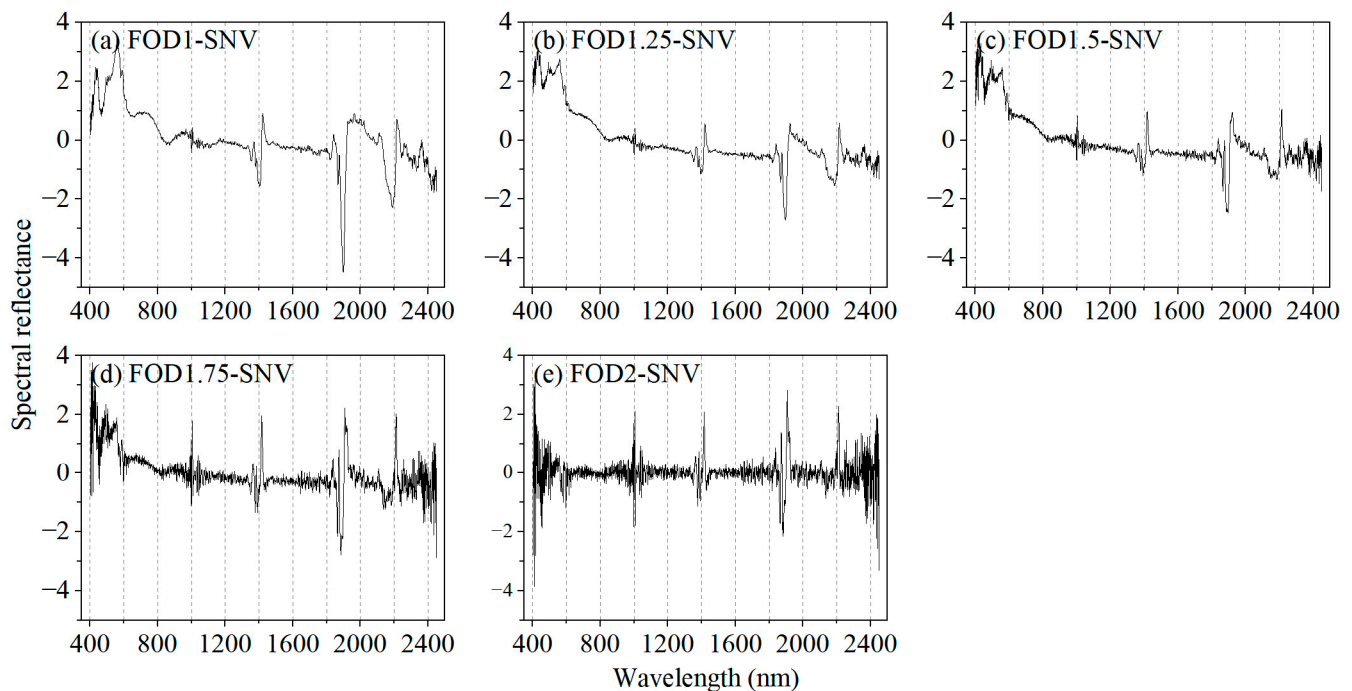


Figure 3. Characterization of spectral curves with different pre-processing.

3.3. Spectral Feature Wavelength

In Figure 4, the black line indicates the correlation coefficient (r) between the soil Cd content and spectral data, while the blue line indicates the characteristic wavelengths chosen with CARS. It is evident that the correlation between the spectral features and Cd content increases significantly as the derivative order increases, while the range of fluctuation of r diminishes progressively. This demonstrates that the noise is suppressed more effectively, making the correlation analysis steadier and more reliable. In addition, the distribution of the blue lines shows that CARS mainly selected the wavelengths with stronger correlation with soil Cd. The results of CARS feature selection are further quantified in Table 3, where CARS selected 69–139 feature wavelengths for soil Cd from 2050 full-spectrum wavelengths, reducing spectral redundancy. Although FOD1.25-SNV showed improved correlation with FOD1-SNV at some wavelengths, most of the improved wavelengths were not sensitive wavelengths for soil Cd. Therefore, the absolute values of maximum positive r and maximum negative r for the characteristic wavelengths extracted by CARS for FOD1.25-SNV were small, although FOD2.0-SNV reached the maximum positive r at 1996 nm ($r = 0.69$) and showed the maximum negative r at 1243 nm ($r = -0.63$). However, the wavelengths of the maximum positive and negative correlation r are different for the different pre-processed spectra, suggesting that there is a difference in the prominence of the FOD for different wavelength ranges.

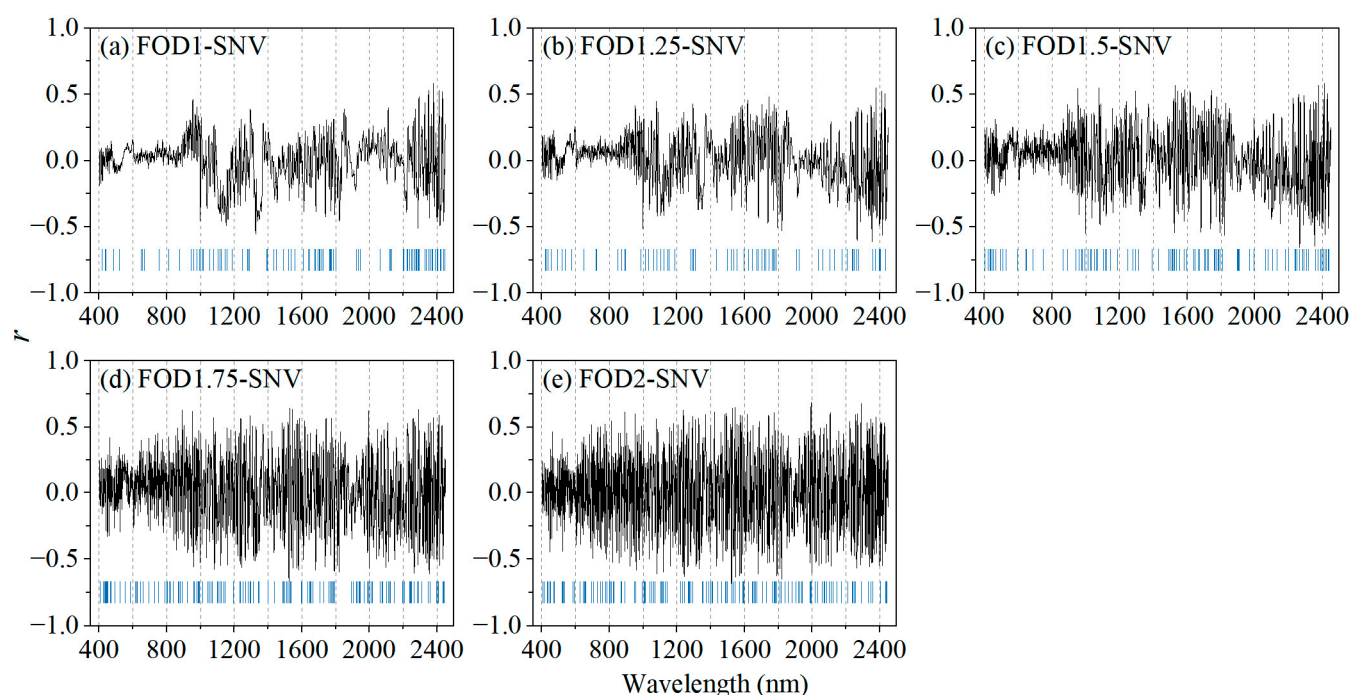


Figure 4. Correlation coefficient r (black line) and CARS feature selection (blue line) for different pre-processed spectra.

Table 3. CARS feature selection for different pre-processing spectra. r , correlation coefficient.

Spectral Pre-Processing	Number of Feature Wavelengths	Maximum Positive r		Maximum Negative r	
		Wavelength	r	Wavelength	r
FOD1-SNV	105	2350	0.53	2287	−0.51
FOD1.25-SNV	69	1295	0.42	1123	−0.41
FOD1.5-SNV	105	1294	0.52	2417	−0.56
FOD1.75-SNV	139	893	0.63	2263	−0.55
FOD2-SNV	139	1996	0.69	1243	−0.63

3.4. Hyperspectral Estimation of Soil Cd Contamination

Table 4 compares the estimation performance of each model on the testing set under different spectral pre-processing. In terms of the impact of various spectral pre-processing methods on the model performance, the estimation accuracy for each model initially increases and then subsequently decreases as the order of the fractional derivative increases. Among them, all models except KNN achieved the highest accuracy under the FOD1.5-SNV pre-processing spectrum. This is mainly because the correlation coefficient between the characteristic wavelength and soil Cd has a significant impact on the modeling accuracy of KNN. In terms of the impact of various models on the estimation performance, the Stacking model outperforms the single machine learning model in most of the pre-processing scenarios, showing its ability to improve the estimation stability by integrating the advantages of multiple models. Under FOD1.5-SNV pre-processing, the Stacking model demonstrated the best estimation ability, with R^2 reaching 0.77, RMSE as low as 0.05 mg/kg, and the RPD value exceeding 2.00. These metrics suggest that the model possesses robust capabilities for quantitative estimation. The results showed that both spectral pre-processing methods and model selection had an impact on the estimation results, and the combination of FOD1.5-SNV and Stacking ensemble learning provided an optimal solution for estimation of soil Cd content.

Table 4. Comparison of the precision of each model testing set with different spectral pre-processing.

Spectral Pre-Processing	Model	R^2	RMSE (mg/kg)	RPD
FOD1-SNV	KNN	0.52	0.07	1.44
	SVR	0.61	0.07	1.60
	RF	0.52	0.07	1.45
	AdaBoost	0.63	0.06	1.64
	XGBoost	0.52	0.07	1.44
	LGBM	0.50	0.07	1.41
	Stacking	0.62	0.06	1.63
FOD1.25-SNV	KNN	0.37	0.08	1.26
	SVR	0.59	0.07	1.55
	RF	0.53	0.07	1.46
	AdaBoost	0.57	0.07	1.52
	XGBoost	0.56	0.07	1.51
	LGBM	0.61	0.06	1.60
	Stacking	0.64	0.06	1.66
FOD1.5-SNV	KNN	0.61	0.07	1.59
	SVR	0.67	0.06	1.75
	RF	0.62	0.06	1.62
	AdaBoost	0.70	0.06	1.84
	XGBoost	0.69	0.06	1.80
	LGBM	0.73	0.05	1.91
	Stacking	0.77	0.05	2.07
FOD1.75-SNV	KNN	0.65	0.06	1.70
	SVR	0.61	0.06	1.61
	RF	0.62	0.06	1.62
	AdaBoost	0.62	0.06	1.63
	XGBoost	0.59	0.07	1.57
	LGBM	0.63	0.06	1.64
	Stacking	0.70	0.06	1.83
FOD2-SNV	KNN	0.63	0.06	1.64
	SVR	0.61	0.06	1.61
	RF	0.62	0.06	1.62
	AdaBoost	0.59	0.07	1.56
	XGBoost	0.60	0.07	1.58
	LGBM	0.58	0.07	1.54
	Stacking	0.66	0.06	1.72

3.5. SHAP Interpretation from Models and Wavelengths

Figure 5 shows the summary and importance plots of the base model SHAP values for optimal Stacking learning. In Figure 5a, the horizontal axis denotes the SHAP value, which quantifies the extent to which the base model contributes to the output of the Stacking model. A positive value signifies that the estimations derived from the base model contribute to an increase in the model output, whereas a negative value indicates that these estimations lead to a decrease in the model output. Each point in the figure corresponds to the estimated value of the base model, with the color gradient transitioning from blue to red signifying a range from low to high estimated values. It can be seen that the SHAP values of the different base models are dispersed on both the positive and negative influence sides. Figure 5b further quantifies the average absolute SHAP value of each model, which is the overall degree of influence of each model on the prediction results. From the figure, SVR, KNN, and AdaBoost have a relatively large degree of influence on the Stacking model.

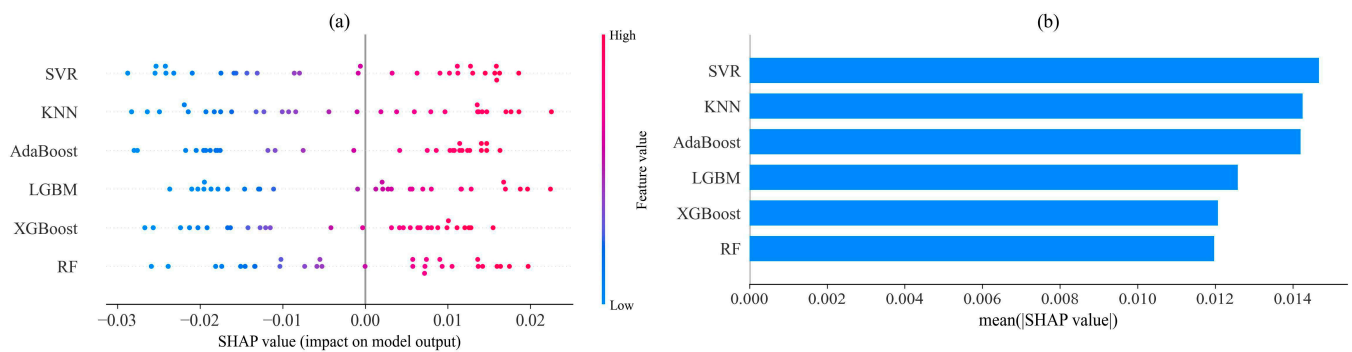


Figure 5. (a) Summary plot and (b) importance plot representing the base model SHAP values in the optimal Stacking model.

Figure 6 shows the summary and importance plots of the SHAP values for the first 10 wavelengths in the best Stacking learning. In Figure 6a, there are differences in the distribution of SHAP values at different wavelengths, and the positive and negative SHAP values of the eigenvalues at the 2417 nm wavelength are larger, indicating that this wavelength has a strong influence on the output in the model. Subsequently, the distribution of SHAP values at each wavelength is gradually concentrated near 0, indicating that the wavelength's impact on the model output is gradually diminishing. Figure 6b further quantifies the average absolute SHAP value of each wavelength, i.e., the overall degree of influence of each wavelength on the prediction results. From the figure, it can be seen that the 2417 nm wavelength contributes most prominently to the model output, and the 10 most important wavelengths affecting the output of the Stacking model are, in order, 2417, 2263, 2395, 1294, 2405, 1525, 1703, 1996, 2292, and 1104 nm.

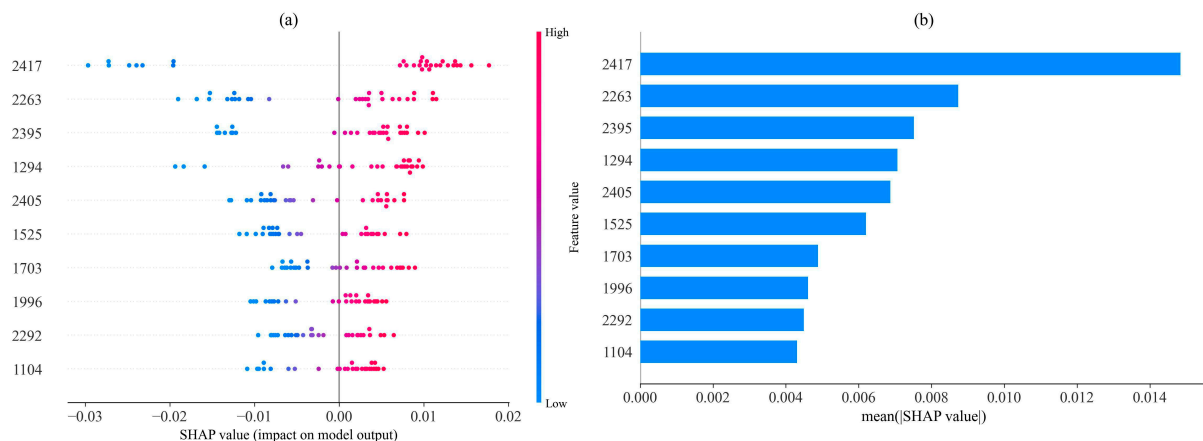


Figure 6. (a) Summary plot and (b) importance plot representing the top 10 wavelength SHAP values in the optimal Stacking model.

4. Discussion

4.1. Effect of Spectral Pre-Processing on Model Performance

In this study, the effects of different FOD combined with SNV spectral pre-processing methods on the performance of soil Cd content estimation models were analyzed by comparing them. The results showed that the FOD1.5-SNV pre-processing method performed most prominently in terms of model accuracy and feature extraction. Comparing the low-order (FOD1 or FOD1.25) and high-order (FOD1.75 or FOD2) derivatives, FOD1.5-SNV was optimal in model performance. It is shown that the appropriate derivative order can effectively enhance the spectral features (e.g., $r = 0.52$ at 1294 nm and $r = -0.56$ at 2417 nm in Table 3) while avoiding the noise interference that may be introduced with the

higher-order derivatives (e.g., the R^2 of the LGBM is reduced to 0.58 under FOD2-SNV in Table 4). This outcome is consistent with the study by Hong et al. [50], who also found that the 1.5-order attained the highest accuracy in their hyperspectral estimation of soil organic matter. This is primarily attributable to the fact that FOD eliminates the baseline drift effect and reduces the overlapping spectral bands. Utilizing smaller intervals for variation can facilitate gradual changes in the spectral information, thereby allowing for the extraction of more nuanced characteristics of the spectral signals [51,52]. Meanwhile, in light of the redundancy and collinearity inherent in hyperspectral data, it is typically necessary to conduct feature selection on the spectral data prior to the modeling process. The CARS algorithm screened 69–139 feature wavelengths of soil Cd in different pre-processed spectra, which only accounted for 3.4–6.8% of the 2050 full-spectrum wavelengths. And the feature wavelengths screened by the CARS algorithm were highly matched with the highly correlated wavelengths. This indicates that the method can effectively identify key spectral information while reducing spectral redundancy, which effectively improves the modeling accuracy and efficiency, consistent with previous studies [53,54].

4.2. The Advantages of the Interpretable Stacking Ensemble Learning Model

Among the modeling results obtained from various spectral pre-processing methods, the Stacking method showed superior estimation performance compared to the base model. Tan et al. [55] also found the Stacking method to be more accurate than other machine learning (e.g., PLSR, SVR, RF, XGBoost, AdaBoost, etc.) methods for estimating heavy metal As, Cr, Pb, and Zn contamination of soils in a mining area by using airborne hyperspectral imagery methods with higher accuracy. Yang et al. [56] utilized Stacking ensemble learning consisting of PLSR, SVR, BPNN, XGBoost, and RF in the high Cu content of farmland soil spectral estimation. Their findings also show that stacked models outperform individual machine learning models. This is primarily because the Stacking ensemble learning model optimizes the base model estimation results through the linear weighting of the meta-model (PLSR). The approach successfully combines the robust parsing capabilities of tree-based models, such as LightGBM and AdaBoost, with the adaptive strengths of SVR in addressing nonlinear relationships. This integration enhances both estimation accuracy and generalization performance [29]. Interestingly, although KNN performs poorly in standalone modeling, it achieves a high contribution in stacking. This is likely because the estimation results of KNN are more suitable for further learning and optimization by the meta-model.

The SHAP method further reveals the transparency of the Stacking model decisions by facilitating the visualization of the contributions made by each base model and wavelength to the ultimate prediction outcomes. For example, SVR did not achieve the best accuracy in the base model estimation (Table 4) but contributed the most in the Stacking model (Figure 5). This shows that the Stacking model can capture complementary information between base models through secondary learning to reduce the bias of the final model. We choose to display the top 10 most important wavelengths. On the one hand, the contribution ratio of each of these 10 wavelengths to all characteristic wavelengths exceeds 2%, and their total contribution is close to 40%. On the other hand, most of the other important feature wavelengths are distributed near the 10 most important wavelengths because, usually, the wavelengths near the important feature wavelengths also have large contributions [32]. Among the top 10 feature wavelengths with the largest contribution, the 2417 nm and 1294 nm wavelengths are ranked 1st and 4th, respectively (Figure 6), which is consistent with the highly correlated wavelengths under FOD1.5-SNV in Table 3. This further explains the sensitivity of the Stacking model to key spectral features that are highly correlated with the characteristic wavelengths of organic matter, clay minerals, and iron oxides [57–59].

Therefore, the interpretable Stacking ensemble learning model in this study is not only able to improve performance by fusing multiple types of models but can also deeply understand the model decision logic, clarify the role of each base model and feature wavelength, and help model optimization and practical application.

4.3. Limitations and Future Work

Although the interpretable Stacking model is able to integrate the advantages of multiple machine learning base models and performs well in improving estimation accuracy and model interpretation, it also has certain limitations. The model is required to incorporate various base models (KNN, SVR, RF, etc.) during its execution. This integration frequently demands substantial computational resources and time as well as specific hardware capabilities. In this study, stacking ensemble learning required approximately 2.9 to 290 times more computing time than a single model. Meanwhile, optimizing a large number of model hyperparameters (including the parameters of each base model and meta-model) makes the modelling process more complex and time-consuming [60]. This study used the Optuna method to perform global, automatic hyperparameter optimization through iteration, effectively improving search efficiency [61]. To reduce the model complexity and increase the modeling efficiency, future research could examine ways to optimize the base learner's selection and combination strategy or simplify the model structure while preserving model performance [56]. In addition, this study mainly considers the effect of FOD on model accuracy. Future studies can further explore more spectral pre-processing methods and their combination methods to further raise the precision and reliability of estimating soil heavy metal pollution [62].

5. Conclusions

In this study, we explored the potential of interpretable Stacking ensemble learning for estimation in soil Cd contamination. We found that the best estimation accuracy came from the FOD1.5-SNV spectral pre-processing with Stacking ensemble learning model ($R^2 = 0.77$, RMSE = 0.05 mg/kg, RPD = 2.07), which has the ability to quantitatively estimate soil Cd contamination. Compared with the modeling accuracy of a single machine learning model in FOD1.5-SNV spectral pre-processing, the Stacking model improved R^2 by 5–26%, reduced RMSE by 0–29%, and improved RPD by 8–30%. This result is mainly attributed to the FOD1.5-SNV pre-processing method, which performs well in removing spectral noise, extracting key features, and enhancing the model's predictive performance. It is the optimal spectral pre-processing method for estimating soil Cd contamination. In addition, the interpretable Stacking model fuses the advantages of multiple base models. Meanwhile, the SHAP method was used to explain how the model made decisions, which demonstrated strong predictive ability and good interpretability. This study shows the advantages of interpretable Stacking ensemble learning in hyperspectral estimation of heavy metal pollution in farmland soils. The findings offer a valuable reference for the future implementation of dynamic monitoring of soil heavy metal contamination in farmland across extensive areas.

Author Contributions: Conceptualization, L.Z. and M.D.; methodology, L.Z. and S.Y.; software, S.Y. and L.Z.; validation, X.X. and S.Y.; formal analysis, X.X. and Z.S.; investigation, L.Z., S.Y. and M.D.; resources, J.L. and Z.S.; data curation, L.Z. and J.L.; writing—original draft preparation, L.Z. and M.D.; writing—review and editing, L.Z., J.L. and Z.S.; visualization, M.D. and X.X.; supervision, J.L. and Z.S.; project administration, L.Z. and J.L.; funding acquisition, L.Z. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (grant number: KYCX24_0221, KYCX25_0249), the National Natural Science Foundation of China (grant number: 324B2060).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We are grateful to the anonymous reviewers for their comments and recommendations.

Conflicts of Interest: Author Meng Ding was employed by the company Jiangsu Environmental Protection Group Suzhou Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Hou, D.; Igalavithana, A.D.; Alessi, D.S.; Luo, J.; Tsang, D.C.; Sparks, D.L.; Yamauchi, Y.; Rinklebe, J.; Ok, Y.S. Metal contamination and bioremediation of agricultural soils for food safety and sustainability. *Nat. Rev. Earth Environ.* **2020**, *1*, 366–381. [\[CrossRef\]](#)
- Khan, S.; Naushad, M.; Lima, E.C.; Zhang, S.; Shaheen, S.M.; Rinklebe, J. Global soil pollution by toxic elements: Current status and future perspectives on the risk assessment and remediation strategies—A review. *J. Hazard. Mater.* **2021**, *417*, 126039. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hou, D.; Jia, X.; Wang, L.; McGrath, S.P.; Zhu, G.; Hu, Q.; Zhao, J.; Bank, M.S.; O'Connor, D.; Nriagu, J. Global soil pollution by toxic metals threatens agriculture and human health. *Science* **2025**, *388*, 316–321. [\[CrossRef\]](#)
- Rashid, A.; Schutte, B.J.; Ulery, A.; Deyholos, M.K.; Sanogo, S.; Lehnhoff, E.A.; Beck, L. Heavy Metal Contamination in Agricultural Soil: Environmental Pollutants Affecting Crop Health. *Agronomy* **2023**, *13*, 1521. [\[CrossRef\]](#)
- Mushtaq, G.; Agrawal, S.; Kushwah, A.; Kumar, A.; Lone, R. Cadmium toxicity in plants and its remediation management: A Review. *Plant Stress* **2025**, *16*, 100894. [\[CrossRef\]](#)
- Zhong, L.; Yang, S.; Chu, X.; Sun, Z.; Li, J. Inversion of heavy metal copper content in soil-wheat systems using hyperspectral techniques and enrichment characteristics. *Sci. Total Environ.* **2024**, *907*, 168104. [\[CrossRef\]](#)
- Shi, T.; Chen, Y.; Liu, Y.; Wu, G. Visible and near-infrared reflectance spectroscopy—An alternative for monitoring soil contamination by heavy metals. *J. Hazard. Mater.* **2014**, *265*, 166–176. [\[CrossRef\]](#)
- Wang, Y.; Zou, B.; Chai, L.; Lin, Z.; Feng, H.; Tang, Y.; Tian, R.; Tu, Y.; Zhang, B.; Zou, H. Monitoring of soil heavy metals based on hyperspectral remote sensing: A review. *Earth-Sci. Rev.* **2024**, *254*, 104814. [\[CrossRef\]](#)
- Zhong, L.; Chu, X.; Qian, J.; Li, J.; Sun, Z. Multi-Scale Stereoscopic Hyperspectral Remote Sensing Estimation of Heavy Metal Contamination in Wheat Soil over a Large Area of Farmland. *Agronomy* **2023**, *13*, 2396. [\[CrossRef\]](#)
- Wang, F.; Gao, J.; Zha, Y. Hyperspectral sensing of heavy metals in soil and vegetation: Feasibility and challenges. *ISPRS J. Photogramm. Remote Sens.* **2018**, *136*, 73–84. [\[CrossRef\]](#)
- Liu, J.; Dong, Z.; Xia, J.; Wang, H.; Meng, T.; Zhang, R.; Han, J.; Wang, N.; Xie, J. Estimation of soil organic matter content based on CARS algorithm coupled with random forest. *Spectrochim. Acta Part A* **2021**, *258*, 119823. [\[CrossRef\]](#)
- Li, H.; Wang, J.; Zhang, J.; Liu, T.; Acquah, G.E.; Yuan, H. Combining Variable Selection and Multiple Linear Regression for Soil Organic Matter and Total Nitrogen Estimation by DRIFT-MIR Spectroscopy. *Agronomy* **2022**, *12*, 638. [\[CrossRef\]](#)
- Ram, B.G.; Oduor, P.; Igathinathane, C.; Howatt, K.; Sun, X. A systematic review of hyperspectral imaging in precision agriculture: Analysis of its current state and future prospects. *Comput. Electron. Agric.* **2024**, *222*, 109037. [\[CrossRef\]](#)
- Liu, P.; Liu, Z.; Hu, Y.; Shi, Z.; Pan, Y.; Wang, L.; Wang, G. Integrating a Hybrid Back Propagation Neural Network and Particle Swarm Optimization for Estimating Soil Heavy Metal Contents Using Hyperspectral Data. *Sustainability* **2018**, *11*, 419. [\[CrossRef\]](#)
- Hong, Y.; Liu, Y.; Chen, Y.; Liu, Y.; Yu, L.; Liu, Y.; Cheng, H. Application of fractional-order derivative in the quantitative estimation of soil organic matter content through visible and near-infrared spectroscopy. *Geoderma* **2019**, *337*, 758–769. [\[CrossRef\]](#)
- Yang, C.; Feng, M.; Song, L.; Jing, B.; Xie, Y.; Wang, C.; Yang, W.; Xiao, L.; Zhang, M.; Song, X. Study on hyperspectral monitoring model of soil total nitrogen content based on fractional-order derivative. *Comput. Electron. Agric.* **2022**, *201*, 107307. [\[CrossRef\]](#)
- Chen, L.; Lai, J.; Tan, K.; Wang, X.; Chen, Y.; Ding, J. Development of a soil heavy metal estimation method based on a spectral index: Combining fractional-order derivative pretreatment and the absorption mechanism. *Sci. Total Environ.* **2022**, *813*, 151882. [\[CrossRef\]](#)
- Cao, J.; Liu, W.; Feng, Y.; Liu, J.; Ni, Y. Predicting nickel concentration in soil using fractional-order derivative and visible-near-infrared spectroscopy indices. *PLoS ONE* **2024**, *19*, e0302420. [\[CrossRef\]](#)
- Zhong, L.; Guo, X.; Xu, Z.; Ding, M. Soil properties: Their prediction and feature extraction from the LUCAS spectral library using deep convolutional neural networks. *Geoderma* **2021**, *402*, 115366. [\[CrossRef\]](#)

20. Wang, X.; Zhang, M.; Zhou, Y.; Wang, L.; Zeng, L.; Cui, Y.; Sun, X. Simultaneous estimation of multiple soil properties from vis-NIR spectra using a multi-gate mixture-of-experts with data augmentation. *Geoderma* **2025**, *453*, 117127. [\[CrossRef\]](#)
21. Barra, I.; Haeefe, S.M.; Sakrabani, R.; Kebede, F. Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances—A review. *Trends Anal. Chem.* **2021**, *135*, 116166. [\[CrossRef\]](#)
22. Qi, C.; Li, K.; Zhou, M.; Zhang, C.; Zheng, X.; Chen, Q.; Hu, T. Leveraging visible-near-infrared spectroscopy and machine learning to detect nickel contamination in soil: Addressing class imbalances for environmental management. *J. Hazard. Mater. Adv.* **2024**, *16*, 100489. [\[CrossRef\]](#)
23. Qi, C.; Zhou, N.; Hu, T.; Wu, M.; Chen, Q.; Wang, H.; Zhang, K.; Lin, Z. Prediction of copper contamination in soil across EU using spectroscopy and machine learning: Handling class imbalance problem. *Smart Agric. Technol.* **2025**, *10*, 100728. [\[CrossRef\]](#)
24. Zhou, W.; Yang, H.; Xie, L.; Li, H.; Huang, L.; Zhao, Y.; Yue, T. Hyperspectral inversion of soil heavy metals in Three-River Source Region based on random forest model. *Catena* **2021**, *202*, 105222. [\[CrossRef\]](#)
25. Belgiu, M.; Drăguț, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [\[CrossRef\]](#)
26. Tsirikoglou, P.; Abraham, S.; Contino, F.; Lacor, C.; Ghorbaniasl, G. A hyperparameters selection technique for support vector regression models. *Appl. Soft Comput.* **2017**, *61*, 139–148. [\[CrossRef\]](#)
27. Zhang, Y.; Liu, J.; Shen, W. A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications. *Appl. Sci.* **2022**, *12*, 8654. [\[CrossRef\]](#)
28. Li, X.; Chen, J.; Chen, Z.; Lan, Y.; Ling, M.; Huang, Q.; Li, H.; Han, X.; Yi, S. Explainable machine learning-based fractional vegetation cover inversion and performance optimization—A case study of an alpine grassland on the Qinghai-Tibet Plateau. *Ecol. Inform.* **2024**, *82*, 102768. [\[CrossRef\]](#)
29. Zou, Z.; Wang, Q.; Wu, Q.; Li, M.; Zhen, J.; Yuan, D.; Zhou, M.; Xu, C.; Wang, Y.; Zhao, Y.; et al. Inversion of heavy metal content in soil using hyperspectral characteristic bands-based machine learning method. *J. Environ. Manag.* **2024**, *355*, 120503. [\[CrossRef\]](#)
30. Dong, H.; Liu, B.; Ye, D.; Liu, G. Interpretability as Approximation: Understanding Black-Box Models by Decision Boundary. *Electronics* **2024**, *13*, 4339. [\[CrossRef\]](#)
31. Viscarra Rossel, R.A.; Behrens, T.; Ben-Dor, E.; Chabrilat, S.; Melo Demattê, J.A.; Ge, Y.; Gomez, C.; Guerrero, C.; Peng, Y.; Ramirez-Lopez, L.; et al. Diffuse reflectance spectroscopy for estimating soil properties: A technology for the 21st century. *Eur. J. Soil Sci.* **2022**, *73*, e13271. [\[CrossRef\]](#)
32. Zhong, L.; Guo, X.; Ding, M.; Ye, Y.; Jiang, Y.; Zhu, Q.; Li, J. SHAP values accurately explain the difference in modeling accuracy of convolution neural network between soil full-spectrum and feature-spectrum. *Comput. Electron. Agric.* **2024**, *217*, 108627. [\[CrossRef\]](#)
33. Li, W.; Jiang, Y.; Ye, Y.; Guo, X.; Shi, Z. Spatiotemporal interpretable mapping framework for soil heavy metals. *J. Clean. Prod.* **2024**, *468*, 143101. [\[CrossRef\]](#)
34. Le, S.; Duan, Y. Determination of Heavy Metal Elements in Soil by ICP-MS. *Chin. J. Inorg. Anal. Chem.* **2015**, *5*, 16–19.
35. Benkhettou, N.; da Cruz, A.; Torres, D.F.M. A fractional calculus on arbitrary time scales: Fractional differentiation and fractional integration. *Signal Process.* **2015**, *107*, 230–237. [\[CrossRef\]](#)
36. Xing, Z.; Du, C.; Shen, Y.; Ma, F.; Zhou, J. A method combining FTIR-ATR and Raman spectroscopy to determine soil organic matter: Improvement of prediction accuracy using competitive adaptive reweighted sampling (CARS). *Comput. Electron. Agric.* **2021**, *191*, 106549. [\[CrossRef\]](#)
37. Li, H.; Liang, Y.; Xu, Q.; Cao, D. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* **2009**, *648*, 77–84. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [\[CrossRef\]](#)
39. Geladi, P.; Kowalski, B.R. Partial least-squares regression: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17. [\[CrossRef\]](#)
40. Keller, J.M.; Gray, M.R.; Givens, J.A. A fuzzy K-nearest neighbor algorithm. *IEEE Trans. Syst. Man Cybern.* **1985**, *SMC-15*, 580–585. [\[CrossRef\]](#)
41. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
42. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
43. 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 Sens. Environ.* **2008**, *112*, 2999–3011. [\[CrossRef\]](#)
44. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
45. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the Advances in Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 3149–3157.

46. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631.
47. Williams, P.; Manley, M.; Antoniszyn, J. *Near Infrared Technology: Getting the Best Out of Light*; African Sun Media: Stellenbosch, South Africa, 2019.
48. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
49. Li, Z. Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Comput. Environ. Urban Syst.* **2022**, *96*, 101845. [\[CrossRef\]](#)
50. Hong, Y.; Chen, Y.; Yu, L.; Liu, Y.; Liu, Y.; Zhang, Y.; Liu, Y.; Cheng, H. Combining Fractional Order Derivative and Spectral Variable Selection for Organic Matter Estimation of Homogeneous Soil Samples by VIS–NIR Spectroscopy. *Remote Sens.* **2018**, *10*, 479. [\[CrossRef\]](#)
51. Schmitt, J.M. Fractional derivative analysis of diffuse reflectance spectra. *Appl. Spectrosc.* **1998**, *52*, 840–846. [\[CrossRef\]](#)
52. Tian, R.; Zou, B.; Li, S.; Dai, L.; Zhang, B.; Wang, Y.; Tu, H.; Zhang, J.; Zou, L. A Model Combining Sensitive Vegetation Indices and Fractional-Order Differential Characteristic Bands for SPAD Value Estimation in Cd-Contaminated Rice Leaves. *Agriculture* **2024**, *15*, 311. [\[CrossRef\]](#)
53. Wei, L.; Yuan, Z.; Zhong, Y.; Yang, L.; Hu, X.; Zhang, Y. An Improved Gradient Boosting Regression Tree Estimation Model for Soil Heavy Metal (Arsenic) Pollution Monitoring Using Hyperspectral Remote Sensing. *Appl. Sci.* **2019**, *9*, 1943. [\[CrossRef\]](#)
54. Zhao, M.; Gao, Y.; Lu, Y.; Wang, S. Hyperspectral Modeling of Soil Organic Matter Based on Characteristic Wavelength in East China. *Sustainability* **2022**, *14*, 8455. [\[CrossRef\]](#)
55. Tan, K.; Ma, W.; Chen, L.; Wang, H.; Du, Q.; Du, P.; Yan, B.; Liu, R.; Li, H. Estimating the distribution trend of soil heavy metals in mining area from HyMap airborne hyperspectral imagery based on ensemble learning. *J. Hazard. Mater.* **2021**, *401*, 123288. [\[CrossRef\]](#)
56. Yang, K.; Wu, F.; Guo, H.; Chen, D.; Deng, Y.; Huang, Z.; Han, C.; Chen, Z.; Xiao, R.; Chen, P. Hyperspectral Inversion of Soil Cu Content in Agricultural Land Based on Continuous Wavelet Transform and Stacking Ensemble Learning. *Land* **2024**, *13*, 1810. [\[CrossRef\]](#)
57. Jia, X.; O'Connor, D.; Shi, Z.; Hou, D. VIRS based detection in combination with machine learning for mapping soil pollution. *Environ. Pollut.* **2021**, *268*, 115845. [\[CrossRef\]](#)
58. Rossel, R.A.V.; Behrens, T. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **2010**, *158*, 46–54. [\[CrossRef\]](#)
59. Hong, Y.; Shen, R.; Cheng, H.; Chen, S.; Chen, Y.; Guo, L.; He, J.; Liu, Y.; Yu, L.; Liu, Y. Cadmium concentration estimation in peri-urban agricultural soils: Using reflectance spectroscopy, soil auxiliary information, or a combination of both? *Geoderma* **2019**, *354*, 113875. [\[CrossRef\]](#)
60. Mohammed, A.; Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *J. King Saud Univ.-Comput. Inf. Sci.* **2023**, *35*, 757–774. [\[CrossRef\]](#)
61. Guo, H.; Wu, F.; Yang, K.; Yang, Z.; Chen, Z.; Chen, D.; Xiao, R. Sentinel-2 Multispectral Satellite Remote Sensing Retrieval of Soil Cu Content Changes at Different pH Levels. *Agronomy* **2024**, *14*, 2182. [\[CrossRef\]](#)
62. Jin, H.; Peng, J.; Bi, R.; Tian, H.; Zhu, H.; Ding, H. Comparing Laboratory and Satellite Hyperspectral Predictions of Soil Organic Carbon in Farmland. *Agronomy* **2024**, *14*, 175. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.