remote sensing

MDPI

Article

A Novel Collaborative Subspace and Class Transfer Model-Based Hyperspectral Transfer Learning Framework

Jia Deng [1,†], Jiawen Wang [1,†], Huixia Li [1,†], Bifeng Hu [2,3,4], Junjie Liu [1], Defang Luo [1,*], Jie Peng [1,5,6] and Zhou Shi [4,7]

1 College of Agriculture, Tarim University, Alar 843300, China; 10757242023@stumail.taru.edu.cn (J.D.); 107572023412@stumail.taru.edu.cn (J.W.); 107572024404@stumail.taru.edu.cn (H.L.); 10757231029@stumail.taru.edu.cn (J.L.); pjzky@taru.edu.cn (J.P.)
2 Department of Land Resource Management, School of Public Administration, Jiangxi University of Finance and Economics, Nanchang 330013, China; hubifeng@zju.edu.cn
3 Philosophy and Social Sciences Laboratory of Data Science in Finance and Economics at the Ministry of Education, Jiangxi University of Finance and Economics, Nanchang 330013, China
4 Zhejiang Key Laboratory of Agricultural Remote Sensing and Information Technology, Zhejiang University, Hangzhou 310058, China; shizhou@zju.edu.cn
5 Key Laboratory of Genetic Improvement and Efficient Production for Specialty Crops in Arid Southern Xinjiang of Xinjiang Corps, Alar 843300, China
6 Key Laboratory of Tarim Oasis Agriculture (Tarim University), Ministry of Education, Alar 843300, China
7 College of Environment and Resource Sciences, Zhejiang University, Hangzhou 310058, China
* Correspondence: 120210019@taru.edu.cn; Tel.: +86-155-9976-8501
† These authors contributed equally to this work.

Highlights

What are the main findings?

- A novel hyperspectral transfer learning framework based on subspace partitioning and class transfer modeling is proposed, which successfully addresses the issues of "same spectrum, different materials" and "different spectra, same material" in cross-region hyperspectral remote sensing through unsupervised clustering and class-level transfer models, effectively resolving negative transfer caused by distribution mismatches between source domain (SD) and target domain (TD).
- Experimental results show that the proposed framework improves R^2 by 5.48–47.37%, reduces *RMSE* by 25.96–94.24%, and enhances *RPD* by 35.46–125.49% on the target domain, significantly outperforming traditional single global transfer models. Additionally, applying large-scale soil spectral databases at regional scales greatly enhances their practicality in estimating local and regional soil properties.

What are the implications of the main findings?

- Methodologically, this study introduces a transfer learning paradigm based on subspaces and category models, overcoming the bottleneck of spectral heterogeneity in cross-regional hyperspectral remote sensing.
- Practically, it enhances the applicability of large-scale global soil databases for local and regional soil property estimation, improving the spatial generalizability of remote sensing inversion models and advancing digital soil mapping technologies.

Check for updates

Academic Editors: Mohammadmehdi Saberioon and Akira Iwasaki

Received: 15 September 2026
Revised: 11 November 2026
Accepted: 25 November 2026
Published: 28 November 2026

Copyright: © 2026 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.

Abstract

Hyperspectral soil technology has proven to be an efficient, low-cost, and non-destructive method for characterizing soil properties but is unable to overcome the limitations of global transfer models in practical cross-regional applications. Specifically, it cannot handle the phenomena of "same spectrum, different materials" and "different spectra, same material" as well as the negative transfer caused by large distributional discrepancies between the

source domain (SD) and target domain (TD). This study proposes a subspace- and class-based transfer learning (TL) framework for hyperspectral soil modeling. Using the Global Soil Spectral Library (GSSL) as the SD and the Chinese Soil Spectral Library (CSSL) as the TD, with soil organic matter (SOM) as the target property, we demonstrate the efficacy of this framework. The results indicate that unsupervised clustering via autoencoders (AECs) achieves high-quality partitioning of the GSSL data and assigns all CSSL samples to GSSL-derived clusters with high coverage. This assignment coverage is distinguished from cross-database generalization performance, which is evaluated by the subsequent class transfer prediction. Moreover, the spectral features of samples from the GSSL and CSSL exhibit a high degree of overlap within corresponding subclasses. Furthermore, the proposed TL framework based on subspace and class transfer models significantly outperforms traditional global transfer models: on the CSSL, it improves R^2 by 5.48–47.37%, reduces *RMSE* by 25.96–94.24%, and increases *RPD* by 35.46–125.49%. Additionally, the framework effectively mitigates negative transfer caused by distribution mismatches between the SD and TD, and each class-specific transfer model performs better on its respective subclass than the global model, thereby validating the hypothesis that class-specific transfer models outperform global transfer models. This study provides a novel methodological framework for cross-regional hyperspectral soil detection, expands the application boundaries of transfer learning and large-scale spectral databases in heterogeneous geological settings, and has the potential to drive a paradigm shift in digital soil mapping from global models to class transfer models.

Keywords: hyperspectral data; spectral clustering; global soil spectral library; subspace-based transfer learning; class transfer model

1. Introduction

Soil is an essential natural asset that plays a foundational role in securing global food supplies, modulating the terrestrial carbon cycle, and maintaining vital ecosystem services [1–3]. The precise acquisition of spatial distribution data for soil properties, including organic matter and clay content, constitutes a fundamental prerequisite for the advancement of precision agriculture, environmental monitoring, and climate change adaptation [4–6]. Soil hyperspectral technology, with its advantages of non-destructiveness, speed and low cost, enables the inversion of soil physicochemical properties through soil reflectance spectra. In recent years, it has become a research hotspot in the fields of soil science and remote sensing [7–11]. At the heart of this technology is the application of modeling algorithms to characterize the complex relationship between soil properties and their spectral signatures, whereby predictions are subsequently made [3,7,8]. Soil hyperspectral technology has evolved from early machine learning (ML) to deep learning (DL), and then to ensemble learning (EL), continuously improving the predictive accuracy of models [12–19]. However, owing to the pronounced spatial heterogeneity of soil and substantial regional variations in pedogenic conditions, a "global model" calibrated for a specific area often suffers a marked decline or even an abrupt failure in predictive performance when transferred directly to a different region [3,8,15]. To address this challenge, researchers began introducing transfer learning (TL), attempting to transfer models developed from large-scale soil spectral data to specific regions for prediction [3,20]. Notwithstanding the success of existing studies in establishing cross-regional spectral transfer models, current approaches nevertheless suffer from two key limitations. First, traditional methods typically construct a single global transfer model, expecting a unified spectral–physicochemical mapping relationship to apply to all soil samples [20]. However, soils formed in different regions and under

different soil-forming conditions may exhibit highly similar spectral reflectance patterns, yet their actual physical and chemical properties can differ significantly [17]. Two soil types with significantly different spectral characteristics may actually have highly similar physical and chemical properties [17,20,21]. We define this situation with the phenomena of "same spectrum, different heterogeneity" and "different spectrum, same homogeneity" in soil. Global transfer models attempt to invert all complex relationships using a fixed one-to-one mapping, which tends to generate spurious correlations when dealing with "isomorphic heterogeneity" and fails to effectively exploit underlying patterns in cases of "heterogeneous isomorphism," ultimately leading to reduced model generalization and limited cross-domain transferability. Second, global transfer models may sacrifice some local optimal fitting capability in pursuit of minimizing overall error transfer [20]. To achieve higher overall prediction accuracy, the model has to "average" features across different subregions. For example, it might discard spectral detail patterns that are highly sensitive to a particular area because those patterns do not hold in other regions or may even cause interference [22,23]. The final result often leads to the global model's prediction accuracy on any single subregion being typically lower than that of a local model trained independently and specifically for that subregion [15,20]. Moreover, traditional transfer learning (TL) typically assumes that the source domain (SD) and target domain (TD) have similar distributions. However, in practical applications, when the distributions of SD and TD differ significantly, direct transfer may instead lead to a "negative transfer" phenomenon [24].

In recent years, some researchers have attempted to significantly improve prediction accuracy by grouping samples based on soil type or spectral similarity and then building independent models for each group [25–28]. Shi demonstrated that both soil type-based and spectral similarity-based grouping models can significantly improve prediction performance [25]. Clustering can group soil samples with more uniform spectral characteristics into the same subclass, enabling independent modeling for each subclass and subsequent prediction within the same category. This approach significantly improves accuracy compared to modeling without classification or modeling based solely on soil type [8,25–28]. Based on this, we propose a research question grounded in subspace transfer learning: can clustering algorithms be used to group soil samples from the source domain (SD) into several clusters based on spectral similarity and then build independent "class-specific transfer models" for each group? When a target domain (TD) sample arrives, it is first matched to the most similar cluster, and then the corresponding class transfer model is invoked for prediction. This approach is expected to fundamentally resolve spectral ambiguity-induced mapping confusion—namely, the "same spectrum, different characteristics" and "different spectra, same characteristics" phenomena—while simultaneously circumventing negative transfer caused by pronounced distributional mismatches between the source and target domains.

In summary, drawing upon the Global Soil Spectral Library (GSSL) as the SD, the Chinese Soil Spectral Library (CSSL) as the TD, and soil organic matter (SOM) as the target attribute, this study systematically evaluates the feasibility and transferability of a subspace transfer-based class-specific modeling framework. Its specific objectives are (1) to reveal the spectral clustering grouping characteristics and spatial distribution patterns of GSSL and CSSL soil samples; (2) to evaluate the predictive accuracy of the "class transfer model" established based on GSSL clustering grouping; (3) to assess the transfer potential of evaluation-based transfer models within the target domain (TD); (4) to compare the advantages of this transfer approach with traditional global transfer models; and (5) to explore the application of transfer models based on GSSL in order to assess transfer capacity across China's provinces.

2. Materials and Methods

2.1. Data

Within a transfer learning (TL) framework, this study designates the Global Soil Spectral Library (GSSL) as the source domain (SD) and the Chinese Soil Spectral Library (CSSL) as the target domain (TD). The integrated GSSL consists of 42,517 visible-near-infrared (Vis-NIR) spectral records derived from 927 distinct soil profiles across 58 countries. This dataset is compiled from several internationally authoritative public soil spectral databases, specifically including the International Soil Reference and Information Centre (ISRIC), the ICRAF Soil–Plant Spectral Diagnostic Laboratory, the African Soil Information Service (AfSIS), the Australian Soil Spectral Library (OzSoil), the NRCS Kellogg Soil Survey Laboratory (KSSL), the Brazilian Soil Spectral Library (BSSL), and the European Union's Land Use and Cover Area frame Statistical Survey (LUCAS) soil database [25,29–32]. To ensure transparency of the data sources, the specific sub-databases integrated into the GSSL are explicitly listed, although they are not differentiated by markers in the spatial distribution map. The spatial coverage extends across six continents (Africa, Asia, Europe, North America, Oceania, and South America), guaranteeing high diversity regarding major soil types, pedogenic environments, and land use patterns. CSSL includes 3976 spectral data entries from 22 provinces in China. Figure 1 illustrates the overall spatial distribution of the GSSL and CSSL sample points. Because the sampling points from different sub-databases are heavily clustered and overlapped at the global scale, differentiating them by markers in Figure 1 would severely compromise the visual clarity of the spatial coverage. Therefore, Figure 1 focuses on macro-level spatial representativeness, while the detailed source components are described in the text. For the aforementioned database, the recorded soil organic matter (SOM) was directly measured, and the reported SOM values were used; for records that did not provide SOM but did provide soil organic carbon (SOC), the SOM was calculated using the SOC × 1.724 formula (i.e., the Van Bemmelen coefficient), and then all SOM values were uniformly converted to g kg^{-1}. Records without either SOM or SOC were excluded.

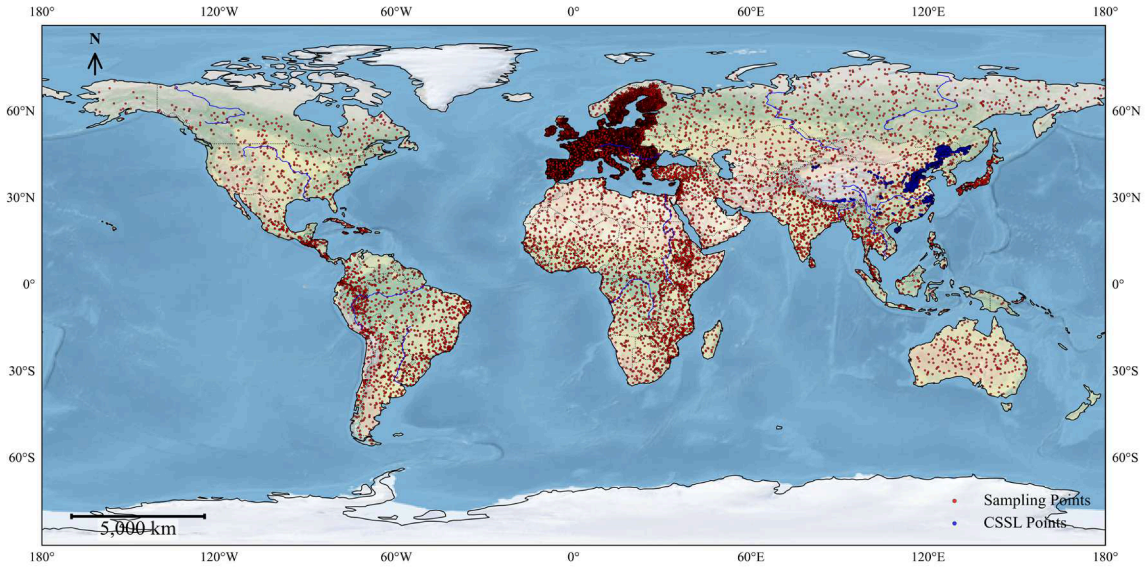

Figure 1. The distribution map of soil sample points for GSSL and CSSL.

It should be noted that although both the GSSL and CSSL datasets provide valuable spectral information, the spectra were acquired using different instruments and measurement settings, and the SOM contents were determined by varying laboratory protocols across different regions. In this study, standard spectral preprocessing (SG smoothing and

SNV) was applied to mathematically reduce baseline shifts and scattering effects. However, a rigorous physical harmonization of the instruments or chemical protocols was not performed prior to modeling, which introduces a degree of uncertainty in the cross-database calibration and validation.

2.2. Spectral Preprocessing and Data Partitioning

To improve the accuracy of soil hyperspectral data, the original spectra from the Global Soil Spectral Library (GSSL) and the Chinese Soil Spectral Library (CSSL) were first preprocessed using Savitzky–Golay (SG) convolution smoothing with a window size of 11 and a polynomial order of 2, followed by Standard Normal Variate (SNV) transformation, to remove noise and scattering effects [33,34]. A total of 42,517 samples from the GSSL were employed to train the Adaptive Embedding Clustering (AEC) model. Considering that the pre-training fine-tuning (PTFT) strategy used in subsequent control experiments requires a certain number of samples for model fine-tuning, and that some provinces in the CSSL have only a small number of samples which is insufficient to allocate adequate fine-tuning and validation samples, this study selected sample data from 11 provinces as demonstration controls for the PTFT method. For each province's dataset, a 7:3 split was performed using the Kennard–Stone (K-S) algorithm. Of these, 70% of the samples were designated as the test set for evaluating the PTFT-established transfer model, with the remaining 30% reserved for fine-tuning the parameters of the pre-trained model. The chosen ratio ensures both an adequate number of samples for fine-tuning the model and a sufficiently large test set to reliably validate the performance of the transferred model [6,19,35]. The statistical characteristics of soil organic matter (SOM) data from the 11 selected provinces are presented in Table 1. A total of 3735 samples are included, with 1122 samples serving as the fine-tuning set. The highest SOM content is found in Xizang, while the lowest is recorded in Xinjiang. The overall mean is 22.22 g kg^{-1}, with a corresponding standard deviation of 7.55 g kg^{-1}, suggesting strong representativeness of the dataset. To rigorously rule out information leakage and avoid optimistic performance estimates, strict data isolation was enforced throughout the entire experimental protocol. The 70% held-out test set of each province was exclusively reserved for the final performance evaluation of the models and was never involved in any feature selection, clustering, hyperparameter tuning, or early stopping processes.

Table 1. CSSL migration of the SOM (g kg^{-1}) data statistics for a certain province.

Province	Samples	Fine-Tuning	Max	Min	Mean	SD
Xinjiang	917	276	80.03	0.95	15.05	9.90
Hebei	122	36	25.72	4.43	13.14	5.87
Henan	313	94	38.27	3.95	15.43	5.68
Heilongjiang	393	118	94.55	7.92	35.73	7.77
Jilin	79	24	35.20	4.31	17.35	7.80
Liaoning	140	42	67.08	3.68	15.00	7.20
Shandong	113	34	27.38	5.53	12.63	5.24
Sichuan	134	40	44.1	9.40	25.55	7.18
Xizang	89	27	194.94	6.8	44.77	9.5
Yunnan	71	21	77.94	4.35	26.80	6.57
Zhejiang	1364	410	78.1	3.71	23.01	10.31
Total	3735	1122	194.94	0.95	22.22	7.55

2.3. The Algorithm to Be Used

2.3.1. Autoencoder Clustering

This study employs autoencoder clustering (AEC) as the clustering algorithm required within the subspace-based and class transfer model transfer framework. The choice of this algorithm is based on our prior systematic evaluation of various traditional and deep clustering algorithms in the field of soil hyperspectral data, which demonstrated that AEC is the optimal method [19]. AEC functions as an unsupervised neural network that operates by compressing input data into a low-dimensional latent space through an encoder before reconstructing it with a decoder, thereby learning compact feature representations and facilitating high-precision unlabeled classification [36,37]. The core of AEC lies in introducing a Self-Expressive Layer and a Collaborative Learning strategy to enhance clustering performance. Its network structure is shown in Figure 2. The AEC is composed of multiple fully connected layers (FCs) stacked together, with the number of neurons decreasing progressively across layers. ReLU activation functions are used to extract deep features from soil spectra. The encoder's output is connected to a self-expressive layer, which learns a coefficient matrix to model the self-expressive relationships among samples. The regularization coefficient λ is set to 0.01 to balance reconstruction error and self-expressive loss. The output from the self-expressive layer is fed into the decoder on one hand; the decoder has a structure symmetric to the encoder and is used to reconstruct the original spectral input, with reconstruction loss calculated using mean squared error (MSE). On the other hand, the output of the self-expressive layer is mapped through another set of fully connected layers and then passed through a Softmax function. The learning mechanism simultaneously optimizes the reconstruction loss, self-expressive loss, and clustering loss, where the clustering loss employs Kullback–Leibler divergence (KL divergence) to align the target distribution with the soft assignment distribution. During training, the batch size is set to 256, the initial learning rate is 0.001, the Adam optimizer is used, and the number of iterations is 200. The detailed architecture and explicit data flow are illustrated in Figure 2. Specifically, the spectral data first flows into the encoder, which consists of multiple fully connected layers, to extract deep features. The encoder's output is then fed into the self-expressive layer to learn a coefficient matrix. From this layer, the data flow bifurcates into two distinct paths: (1) the primary path flows into the decoder to reconstruct the original spectral input, where the reconstruction loss is calculated using mean squared error (MSE); (2) the secondary path is mapped through another set of fully connected layers and a Softmax function to output the soft assignment distribution, where the clustering loss is calculated using Kullback–Leibler (KL) divergence. By simultaneously optimizing these two flows along with the self-expressive loss, the network achieves high-precision clustering. For the specific details of AEC, please refer to the article published by Deng et al. in 2026 [38]. To prevent target-domain information from leaking into the feature space construction, the AEC model was trained strictly using the unlabeled source domain (GSSL) spectral data. No data from the target domain (CSSL), including both the fine-tuning subset and the test subset, was utilized during the latent subspace construction.

2.3.2. Genetic Algorithm

The genetic algorithm (GA) is a global optimization search method that simulates natural selection and genetic mechanisms, iteratively evolving populations through operations such as selection, crossover, and mutation, making it suitable for feature band selection in hyperspectral data [6,39]. Numerous studies have shown that the GA can effectively eliminate redundant and noisy variables in soil hyperspectral analysis, retaining the feature bands most closely related to soil properties [6,19,40]. This study employed the GA to perform spectral feature selection on the sample data of GSSL, CSSL, and the selected

11 provinces, with specific parameter settings detailed in the literature [6,19,40]. After optimization and selection by the GA, the extracted feature spectral bands are mainly concentrated around 550 nm, 900 nm, 1100 nm, 1350 nm, 1800 nm, and 1900 nm, These bands are highly consistent with the spectral response characteristic regions of SOM confirmed by Ben-Dor et al. [41]. To strictly avoid feature selection bias, the genetic algorithm (GA) was used exclusively on the source domain (GSSL) data and the designated fine-tuning/training subsets. The 70% held-out test sets were completely excluded from the GA optimization process to ensure a fully unbiased evaluation.

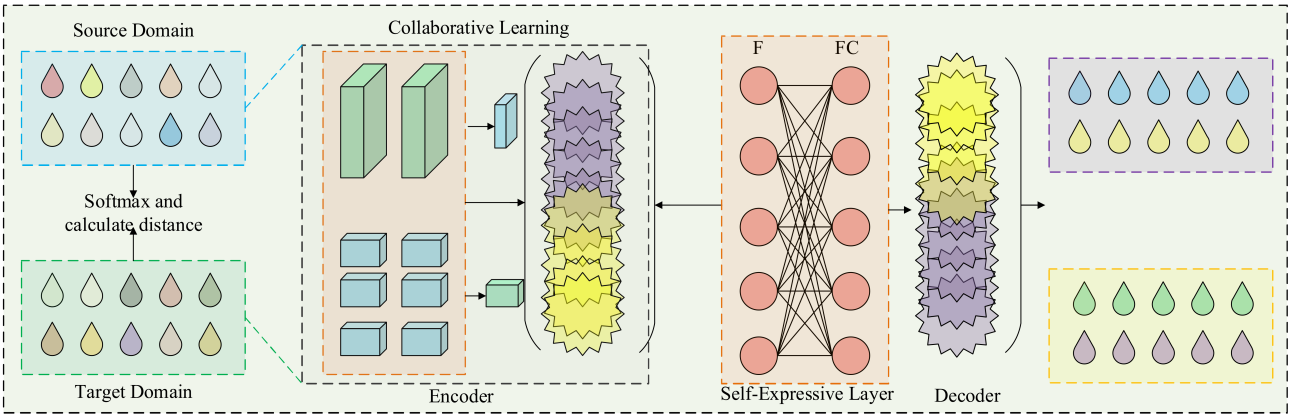

Figure 2. Schematic diagram of the autoencoder clustering (AEC) architecture and its data flow. The network consists of an encoder, a self-expressive layer, and a decoder. The input spectra are compressed into latent features by the encoder, which are then processed by the self-expressive layer. The output diverges into two paths: (1) reconstructing the original spectral input via the decoder (minimizing MSE), and (2) generating clustering assignments via fully connected layers and a Softmax function (minimizing KL divergence).

2.3.3. Principal Component Analysis

Principal component analysis (PCA) operates as a classical unsupervised linear dimensionality reduction technique. It applies an orthogonal transformation to convert the original high-dimensional spectral variables into a reduced number of linearly uncorrelated principal components, which are ranked according to their variance contribution rates in descending order, thereby preserving the principal sources of variability in the original data [6,42,43]. In soil spectroscopy research, beyond its utility in data compression and noise filtering, PCA also enables the intuitive comparison of feature structures across different datasets by projecting all samples into a shared principal component space, thereby revealing similarities or differences in their spectral characteristics [6,23,43]. PCA first performs singular value decomposition on the standardized matrix obtained by merging different datasets, calculates the eigenvalues and eigenvectors of the covariance matrix, and extracts the top principal components whose cumulative variance contribution exceeds 95%. Based on this, each dataset is projected onto a space formed by the first two or three principal components, and two-dimensional or three-dimensional scatter plots are generated. The distribution overlap, clustering trends, and separation among samples from different datasets are then observed [43–45]. This method has become a commonly used approach for qualitatively assessing feature consistency and transfer learning potential among different datasets [46].

2.3.4. CNN

Convolutional neural networks (CNNs) represent a category of deep learning (DL) architectures that have been extensively utilized in the field of spectral analysis [47]. CNNs

can progressively extract hierarchical features from data through convolutional layers, moving from low-level edges and textures to high-level abstract semantics, thereby enhancing the model's ability to understand complex nonlinear relationships in soil spectra [23,47–49]. In soil hyperspectral applications, CNNs have been successfully used to predict various soil properties such as soil organic matter (SOM), moisture, and nitrogen, demonstrating superior fitting capability and generalization performance compared to traditional machine learning (ML) methods [50,51]. This study employs a 1D-CNN to evaluate the subspace transfer and pre-training and fine-tuning (PTFT)-based modeling approach. The CNN architecture comprises 4 convolutional layers with 16, 32, 64, and 128 filters, respectively, each using a kernel size of 3. To extract complex nonlinear features from the soil spectra, ReLU activation functions are applied after each convolutional layer. A max-pooling layer with a window size of 2 follows each convolutional layer to reduce feature dimensionality while retaining key information. After the convolutional blocks, the extracted feature maps are flattened and passed through fully connected layers. Specifically, the final regression output layer contains a single neuron with a linear activation function to directly predict the continuous SOM values without restricting the output range. The model is trained with the Adam optimizer, and mean squared error (MSE) serves as the loss function to minimize prediction error. All algorithms are implemented in Python 3.10.0, and the 1D-CNN is built using the TensorFlow library.

2.3.5. Pre-Training and Fine-Tuning

Pre-training and fine-tuning (PTFT) algorithms, as a model-based transfer learning (TL) approach, have been widely applied in fields such as Artificial Intelligence (AI) and visual communication [52]. This study selects the combination of convolutional neural networks (CNN) and PTFT as a baseline method to evaluate the transfer potential of the proposed subspace- and class-model-based TL framework. Following the PTFT procedure, CNN is first pre-trained on source domain (SD) data to enable the model to extract representative feature representations [53,54]. Subsequently, the pre-trained network architecture is appropriately adjusted, and certain layers are frozen to preserve the learned feature relationships and dependency information [53]. After entering the fine-tuning phase, only the parameters of the lower-level network are updated initially, while the weights of the other pre-trained layers remain fixed. Once the model's performance on the target domain (TD) dataset reaches a certain level, the deeper layers are gradually unfrozen, followed by global fine-tuning of the entire model [52]. To ensure a rigorous validation protocol, the 30% fine-tuning subset was strictly isolated from the 70% held-out test set. Any internal validation required for early stopping or hyperparameter tuning during the fine-tuning phase was performed by further splitting the 30% fine-tuning subset (e.g., using a 5-fold cross-validation internally). The 70% test set remained entirely unseen until the final evaluation.

2.4. Transfer Learning Framework Based on Subspaces and Class Transfer Models

Subspace-based transfer learning (TL) methods aim to project or align source domain (SD) and target domain (TD) data into a shared low-dimensional subspace based on feature distribution similarity, thereby reducing or aligning the distribution differences between the two domains and enhancing the effectiveness of cross-domain knowledge transfer [53,54]. This study proposes a class-based transfer strategy within this framework. The overall workflow is illustrated in Figure 3 and consists of three sequential stages:

Stage 1 comprises source domain clustering. First, the framework utilizes all 42,517 unlabeled soil spectral samples from the Global Soil Spectral Library (GSSL) and applies an autoencoder (AEC) algorithm for unsupervised clustering. Based on the intrin-

sic similarity of spectral features, the data are divided into N clusters, with each cluster containing samples exhibiting similar spectral response patterns. After clustering, the trained clustering model and group centroids are saved. Stage 2 comprises class-specific model training. Next, for each of the N clusters, a dedicated convolutional neural network (CNN) inversion model is independently trained. For a specific cluster, the spectral data and their corresponding soil organic matter (SOM) labels within that cluster are used as the input features and target values, respectively. The CNN learns a specific regression mapping from the spectral data to the SOM values, with its own unique set of learnable weights. During the training procedure, these weights are optimized independently by minimizing the mean squared error (MSE) loss. The MSE loss is calculated as the average of the squared differences between the predicted SOM values and the actual labels across all samples within that cluster. Due to the high spectral consistency among samples within each cluster, these class-specific models can more accurately capture localized spectral–attribute relationships without interference from other clusters [19]. Stage 3 comprises target domain matching and prediction. Finally, in the transfer prediction stage, target samples from the Chinese Soil Spectral Library (CSSL) are input into the saved clustering model. The model matches each sample to the most similar spectral group based on the similarity between the sample's spectral features and the existing group centroids, assigning each CSSL sample to its closest spectral group. Subsequently, the pre-trained CNN model corresponding to that group is invoked to predict the SOM content of the CSSL samples. During the target-domain matching and prediction stage, the parameters of both the AEC encoder and the class-specific CNN models were strictly frozen. No parameter updates or gradient descents were performed using the target test samples, ensuring that the reported performance reflects the true zero-shot transfer capability of the models.

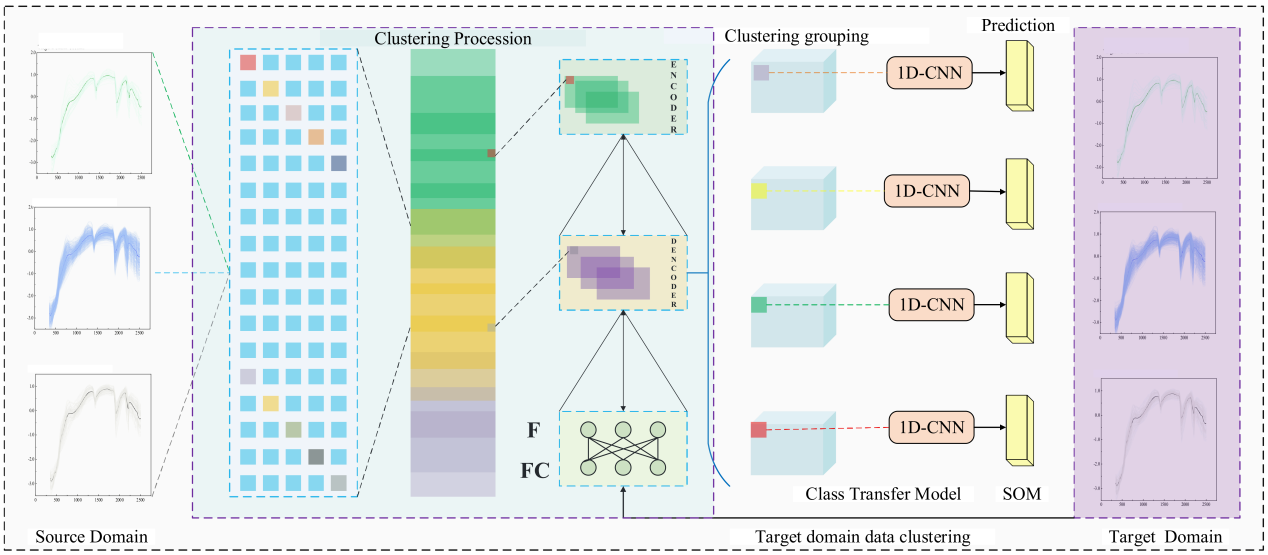

Figure 3. The overall workflow of the proposed subspace and class-model-based transfer learning framework. The process is divided into three stages. Stage 1: GSSL source domain spectra are clustered into N groups using the AEC model. Stage 2: For each cluster, a dedicated 1D-CNN model is trained using the clustered spectra and their corresponding SOM labels. Stage 3: CSSL target domain spectra are matched to the most similar cluster centers, and the corresponding pre-trained 1D-CNN is invoked to predict SOM.

To formally and mathematically clarify the relationship between the learned latent space, the cluster representations, and the target-domain assignment procedure, the framework logically defines these concepts as follows. First, the source-domain spectral data is processed by the AEC encoder, which acts as a mathematical mapping function to com-

press the original high-dimensional spectra into a low-dimensional latent feature space. This learned latent feature space is exactly what our framework refers to as the "spectral subspace". Within this latent subspace, the clustering algorithm groups the source-domain samples into N clusters, and the geometric centers of these clusters are defined as the cluster centroids. These centroids serve as the concrete mathematical representations of the clusters within the shared subspace.

During the target-domain assignment stage, each target-domain sample is projected into this exact same latent subspace using the identical, frozen encoder, yielding its latent feature representation. The assignment procedure is mathematically defined as a deterministic nearest-neighbor search: the framework calculates the distance (e.g., Euclidean distance) between the target sample's latent features and all existing cluster centroids in the subspace and strictly assigns the sample to the cluster with the minimum distance. Once the target sample is assigned to this optimal cluster, the corresponding class-specific transfer model is dynamically invoked to predict the target property (SOM). This explicit formulation ensures that the subspace, the cluster representations, and the target assignment are mathematically unified within a single, rigorous feature space, providing a solid theoretical foundation for the proposed transfer learning framework.

2.5. Evaluation Indicators

The silhouette coefficient is used in this study to assess clustering quality. It quantifies two aspects for each sample: the similarity to its own cluster and the dissimilarity to the nearest other cluster. Ranging from −1 to 1, higher silhouette values signify more distinct cluster structures, stronger intra-group cohesion, and greater inter-group separation [55]. In addition, the silhouette coefficient is commonly used to evaluate the clustering quality of a clustering model on unlabeled spectral data, providing a basis for determining the optimal number of clusters [56,57].

On the other hand, to comprehensively evaluate the predictive performance and transfer potential of models established under different transfer frameworks, this study employs multiple quantitative metrics for assessment. The coefficient of determination (R^2) is used to measure the extent to which the model explains the target variable, with values closer to 1 indicating better fit [58]. Root mean square error ($RMSE$) reflects the overall deviation between predicted and actual values; the smaller the $RMSE$, the higher the prediction accuracy [23,49]. The relative prediction difference (RPD) is defined as the ratio of the standard deviation of the test set samples to the $RMSE$ and is used to evaluate a model's ability to predict soil property variability. Generally, an RPD greater than 2.0 indicates strong quantitative predictive capability, values between 1.4 and 2.0 are considered acceptable, and values below 1.4 suggest poor predictive performance [23,50].

3. Results

3.1. GSSL Clustering Results

To evaluate the unsupervised clustering capability of the autoencoder (AEC) algorithm on soil spectral data, this study input all 42,517 samples from the Global Soil Spectral Library (GSSL) into the model for clustering. Using the silhouette coefficient as the evaluation metric for cluster number, a search was conducted within the candidate range of [2–30]. The results showed that the silhouette coefficient reached its maximum value (0.772) when the number of clusters N = 18, indicating an optimal balance between intra-cluster compactness and inter-cluster separation at this point. Therefore, the GSSL samples were ultimately divided into 18 spectrally similar groups. To visually present the clustering results, this study performed principal component analysis on the spectra of samples within each group, compressing the high-dimensional spectral features into the first two principal components

(PC1 and PC2), and plotted a two-dimensional scatter distribution map for all samples, as shown in Figure 4. The figure clearly shows that samples from different groups exhibit distinct clustering patterns on the PC1-PC2 plane, with each of the 18 groups occupying relatively independent regions. The boundaries between groups are well defined, with only minor overlapping areas, fully demonstrating that the AEC algorithm can effectively separate samples based on spectral feature similarities. Within the internal structure of each group, the majority of sample points are tightly clustered around the center, exhibiting low dispersion, thus indicating a high degree of consistency in spectral characteristics among samples within the same group. For example, the sample clusters in Group 3, Group 7, and Group 12 are compact with small radii, indicating highly homogeneous soil spectral patterns within these groups. Although Groups 1 and 5 have larger sample sizes, they still maintain good cohesion without showing obvious subgroups or dispersion trends. A few groups show modest elongation along the PC1 axis; however, the absence of significant inter-group overlap suggests that the intra-group natural variation, though present, does not compromise the overall clustering integrity. Overall, the 18 groups showed good inter-group separation and high intra-group consistency, indicating that AEC can partition GSSL samples into spectrally distinct groups using only spectral information, providing a grouping foundation for subsequent development of "transfer learning models".

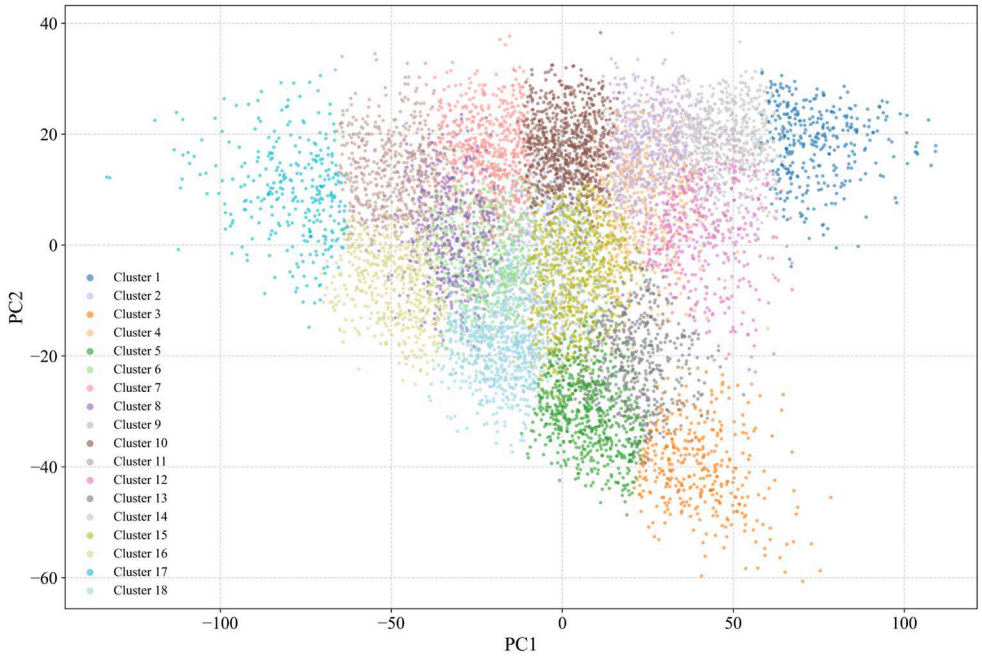

Figure 4. The dimensionality-reduced spectral feature distribution of 18 clustering groups.

Table 2 presents the descriptive statistics of soil organic matter (SOM) content for 18 within-group samples. In terms of sample size, there is significant variation among the groups, with Group 4 having the largest number of samples (5163) and Group 6 the smallest (198). There is a certain range between the maximum and minimum SOM contents across all groups. For example, Group 11 has a minimum SOM value of 0.87 g kg^{-1} and a maximum of 124.77 g kg^{-1}, with a standard deviation of 7.55 g kg^{-1}, indicating a large range. In contrast, Groups 1 and 13 have mean values around 32–33 g kg^{-1}, with a standard deviation of approximately 11 g kg^{-1}, showing relatively higher variability. This result further confirms the presence of "same spectrum, different materials" and "different materials, same spectrum" phenomena in practical applications of soil hyperspectral data.

Table 2. The samples after clustering and their descriptive statistics based on SOM (g kg^{-1}).

Group	Samples	Max	Min	Mean	SD
1	4668	129.62	5.68	32.93	11.79
2	978	63.88	8.99	32.6	10.98
3	1432	45.75	12.07	12.88	6.74
4	5163	83.65	2.69	17.22	9.19
5	2432	103.31	4.4	19.11	10.78
6	198	36.05	12.06	23.38	4.8
7	563	29.1	6.82	7.25	4.46
8	4785	133.28	2.24	13.33	9.56
9	1032	35.36	1.03	17.19	7.54
10	769	47.63	2.6	11.27	9.01
11	4066	124.77	0.87	7.16	7.55
12	3284	82.37	4.7	23.84	8.95
13	291	115.09	9.96	32.66	11.03
14	1864	47.46	4.19	9.05	8.65
15	2497	101.07	2.7	22.5	9.68
16	2515	108.69	6.05	27.28	6.52
17	1210	90.87	1.2	43.43	7.56
18	2253	135	4.74	16.99	9.05

3.2. Establishing a Class Transfer Model

The R^2, *RMSE*, and *RPD* values of the CNN class transfer models built separately for 18 groups are summarized in Figure 5, with results averaged after 10-fold cross-validation. Overall, the R^2 values of all models ranged from 0.71 to 0.85, with an average R^2 of 0.79 across the 18 models, indicating that the local models established within each group had good fitting performance. The minimum *RMSE* was 1.92 g kg^{-1}, the maximum was 3.66 g kg^{-1}, and the average *RMSE* was 2.73 g kg^{-1}, indicating a relatively low overall prediction error. The RPD ranged from 1.38 to 5.40, with Group 1 achieving an *RPD* as high as 5.40, demonstrating excellent predictive performance. Groups 6 and 7 had *RPD* values of 1.38 and 1.43, respectively, slightly below 2.0, which still represents acceptable predictive ability. Overall, most of the classification models achieved a good level of prediction accuracy.

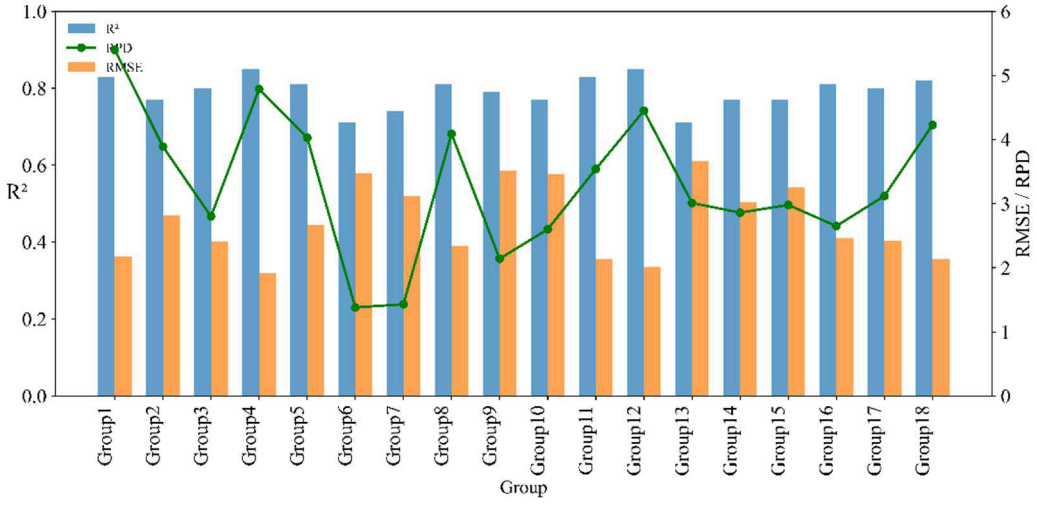

Figure 5. The average R^2, *RMSE* and *RPD* of the 10-fold cross-validation for the 18 group-based class transfer models.

3.3. CSSL Cluster Analysis

The 3735 soil spectral samples from the Chinese Soil Spectral Library (CSSL) were assigned to clusters using a pre-trained autoencoder (AEC) clustering model based on their spectral similarity to existing cluster centers. All CSSL samples were found to be assigned into nine groups (Groups 2, 4, 5, 6, 9, 10, 17, 18). To visually illustrate the allocation results, all CSSL samples were reduced to the first two principal components using PCA, and their group labels were marked with different colors, resulting in the feature distribution plot shown in Figure 6. Overall, the distribution of CSSL samples on the PC1-PC2 plane largely overlaps with the projection area of the GSSL source domain, indicating that AEC effectively classifies the soil spectral feature spaces of the two datasets. Nine groups exhibit distinct clustering patterns in the two-dimensional plane, with most sample points within each group tightly clustered around their respective group centers, indicating low intra-group dispersion and high spectral consistency among CSSL samples within the same group. Notably, all CSSL samples were assigned to a group, with no unmatched outliers observed, indicating that the GSSL-derived cluster centers provide high assignment coverage for CSSL spectral data. However, this coverage reflects the inclusiveness of the clustering model rather than its cross-database generalization capability, which must be assessed through prediction performance (Section 3.4). A few sample points are located in the transitional region between the two groups, which may be due to their spectral characteristics exhibiting intermediate traits of both groups. Overall, the CSSL samples show a well-defined clustering pattern in PCA space, with tight groupings within each class and clear separation between groups, indicating that the AEC model can assign CSSL samples into the GSSL-derived spectral subspaces in a consistent manner based on spectral similarity. This assignment consistency, however, does not by itself constitute evidence of cross-database generalization.

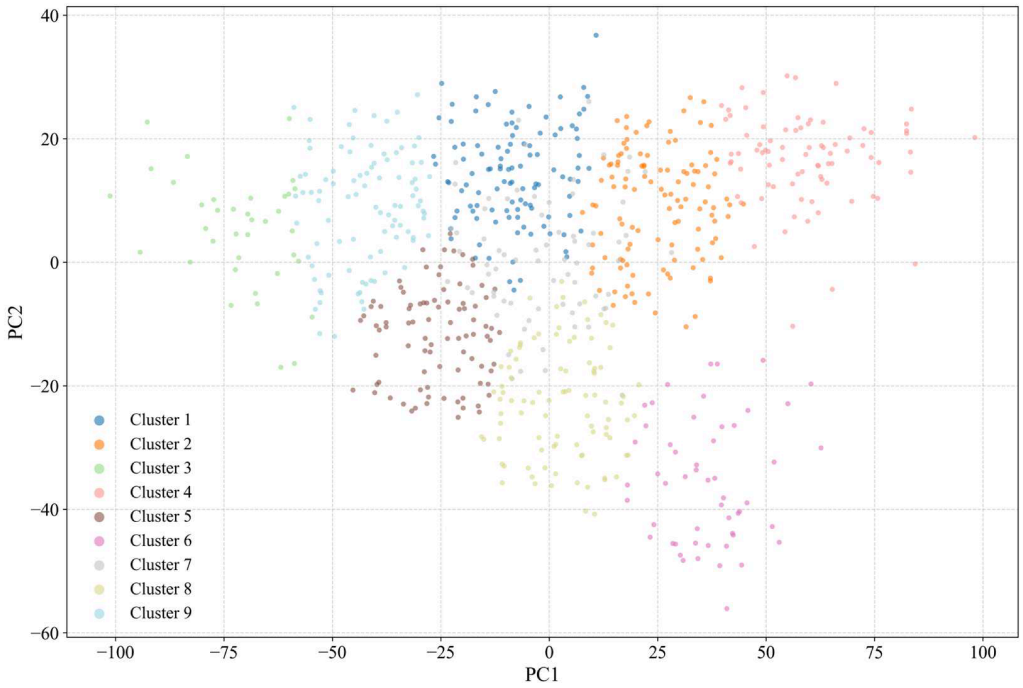

Figure 6. Data visualization after CSSL has been matched by the AEC clustering model.

To further verify the spectral feature consistency between GSSL and CSSL samples within the same group, the above nine groups were selected, and PCA dimensionality reduction visualization was performed separately on the source domain (SD) and target domain (TD) samples within each group. The results are shown in Figure 7. The two-

dimensional distribution plots from each group clearly show that, within every subgroup, GSSL and CSSL samples exhibit highly overlapping patterns in the PCA space, tightly interwoven around a common cluster center, with no obvious separation or displacement observed. For example, in Groups 1 and 4, CSSL samples are almost entirely embedded within the distribution area of GSSL samples; in Groups 8 and 11, although a small number of CSSL samples are slightly scattered at the edges, they remain generally continuous with the blue cluster. In Groups 12 and 15, due to the relatively small number of CSSL samples, their distribution area is completely overlapped by GSSL samples, without forming an independent subpopulation. Groups 2, 5, and 17 also exhibit excellent distribution consistency, with no systematic offset or dispersion trend between red and blue points. The above results indicate that the clustering model trained on GSSL can assign CSSL samples to the GSSL-derived clusters with high coverage, providing a basis for subsequent prediction using the class transfer model. Whether this assignment translates into genuine cross-database generalization is evaluated in Section 3.4.

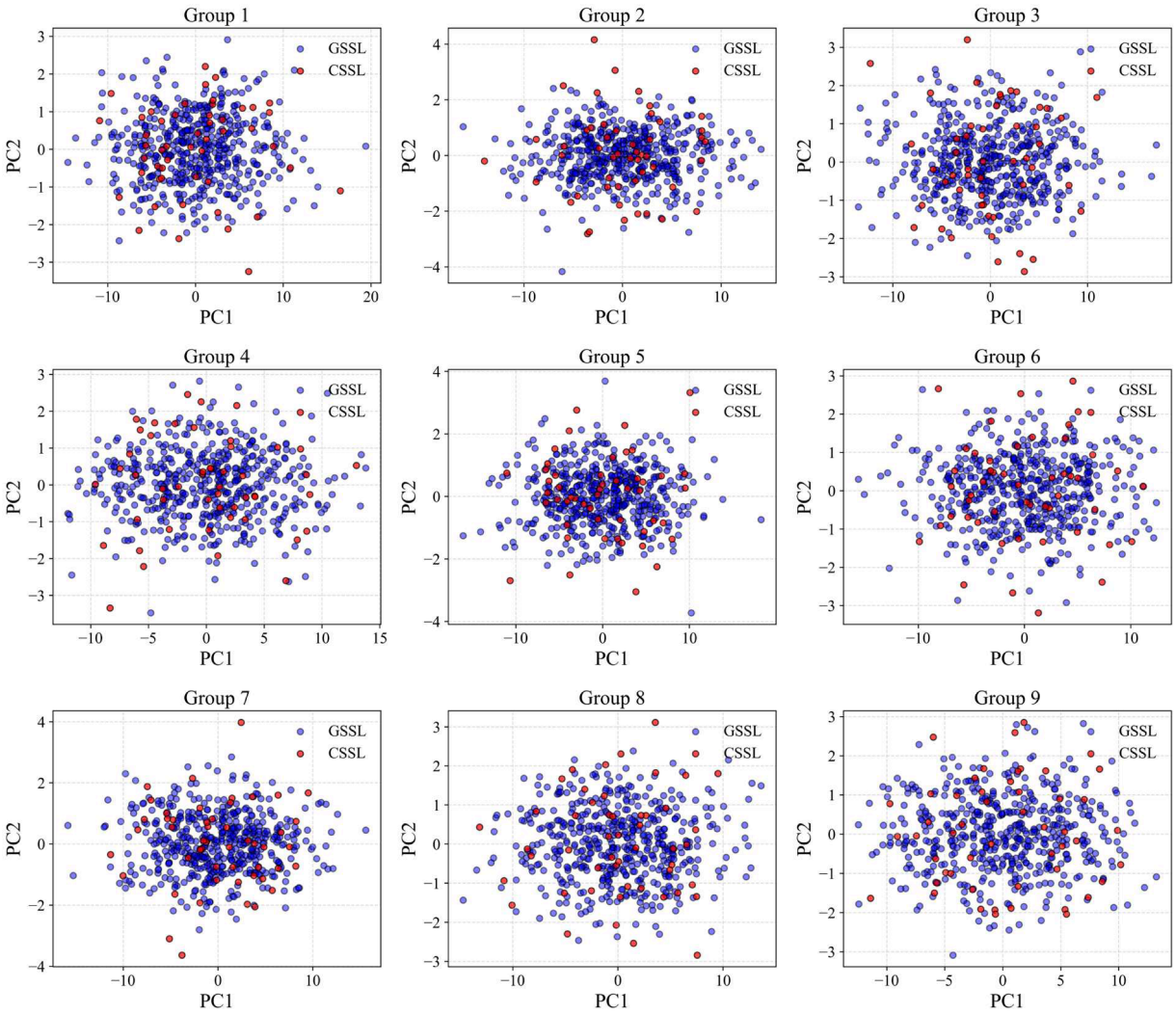

Figure 7. Similarity test of soil spectral characteristics after clustering and grouping of GSSL and CSSL results.

3.4. The Accuracy of the Class Transfer Model When Transferred to CSSL

The samples assigned to nine groups in CSSL were input into the corresponding pre-trained "class transfer models" based on GSSL, and the soil organic matter (SOM) content of each sample in CSSL was predicted. The results are shown in Figure 8. Overall, the

nine class transfer models demonstrated strong predictive performance on the CSSL target domain, with R^2 values ranging from 0.78 to 0.84. Among them, Group 4 had the highest R^2, while Group 8 had the lowest. The average R^2 across all nine models was 0.80, indicating that the transferred models could explain the majority of the SOM variation in the TD samples. The *RMSE* ranged from 2.25 to 3.05 g kg^{-1}, with an average of 2.79 g kg^{-1}. The *RPD* values for the nine transfer models ranged between 2.15 and 2.48. The above results indicate that the model has strong predictive capability. Notably, the data points within each group are closely clustered around the 1:1 diagonal line, with no systematic overestimation or underestimation observed.

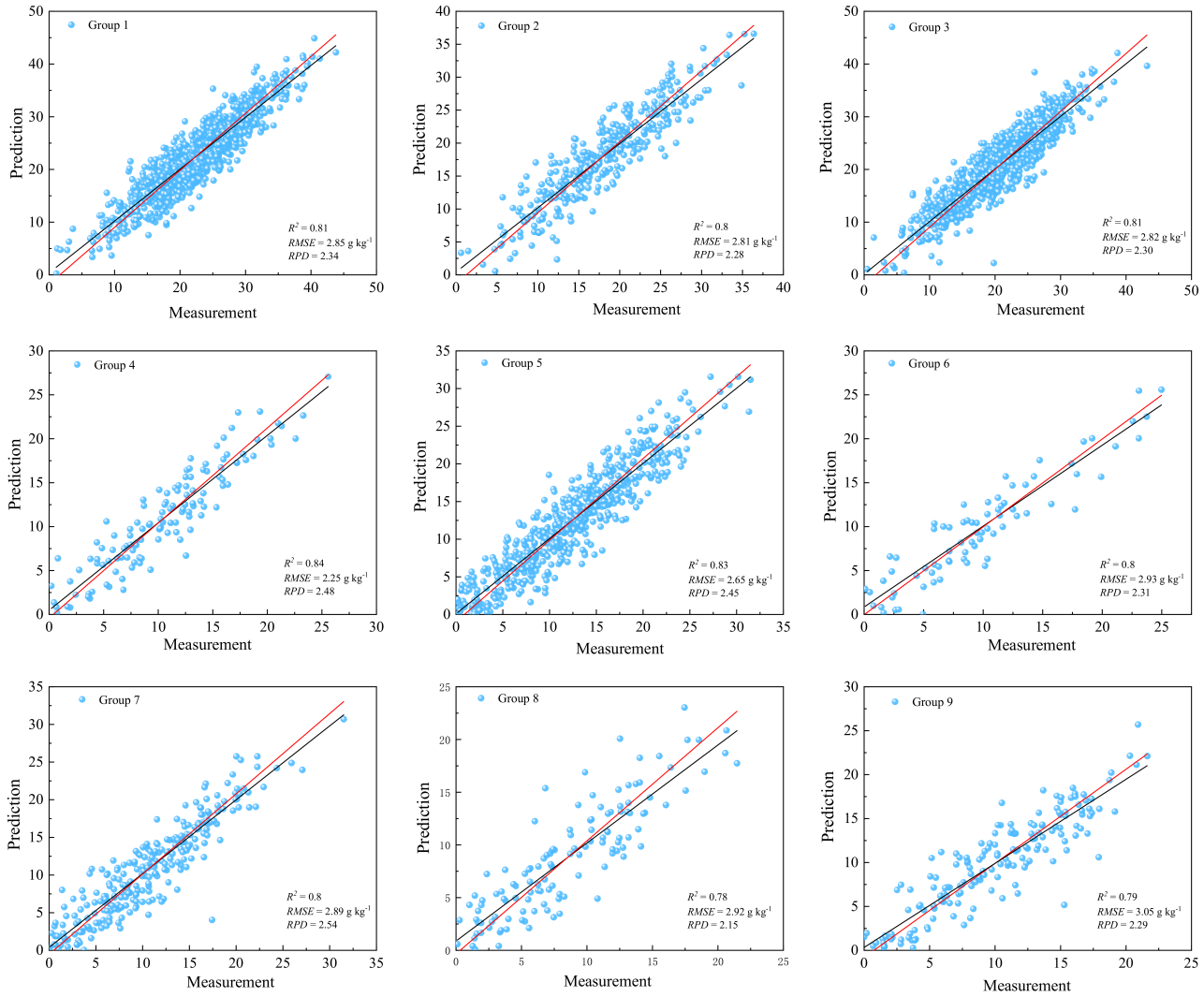

Figure 8. The scatter plot of the actual values and predicted values of the class-specific transfer model of GSSL after being transferred to CSSLtransfer.

3.5. Pre-Training and Fine-Tuning Class-Specific Transfer Models

To further validate the transfer potential of the proposed framework, a CNN-based pre-training and fine-tuning (PTFT) method was selected as the baseline. To ensure a strictly fair evaluation, both methods were assessed on exactly the same 70% held-out test samples for each province. For the PTFT method, the CNN was pre-trained on all 42,517 GSSL samples with 10-fold cross-validation, transferred to each province, and fine-tuned using a 30% target-province subset. For the proposed class transfer model, the provincial test samples were matched to their corresponding GSSL spectral clusters using the pre-trained AEC model, and the corresponding class-specific CNN model was invoked for prediction.

The class transfer model did not use any target-province SOM labels for training, fine-tuning, calibration, or model selection. Because a single province may contain samples matched to multiple spectral clusters, the final performance for each province was obtained by averaging the R^2, RMSE, and RPD values of the matched class-specific models. All reported metrics in Table 3 were computed exclusively on the same 70% held-out test samples for both methods.

Table 3. Comparison table of the performance of subspace and class transfer model transfer models and pre-trained fine-tuning transfer models in the transfer performance of each province in CSSL. Performance comparison between the class transfer model and the PTFT baseline on the **identical 70% held-out test samples** for each province in CSSL. The PTFT model was fine-tuned using a 30% target-province subset, whereas the class transfer model was applied directly without any target-province SOM labels. Because a single province may contain samples matched to multiple clusters, the province-level metrics for the class transfer model are reported as the **mean of the R^2, RMSE, and RPD values of the matched class-specific models**. All metrics were computed on the same 70% test samples for both methods.

Province	Class Transfer Model (Mean of Matched Clusters) R^2	Class Transfer Model (Mean of Matched Clusters) *RMSE*	Class Transfer Model (Mean of Matched Clusters) *RPD*	PTFT Model R^2	PTFT Model *RMSE*	PTFT Model *RPD*
Xinjiang	0.8	2.34	4.23	0.7	4.33	2.29
Hebei	0.81	2.3	2.55	0.65	4.97	1.18
Henan	0.83	2.11	2.69	0.73	3.87	1.47
Heilongjiang	0.82	2.24	3.47	0.68	4.35	1.79
Jilin	0.77	2.41	3.23	0.64	4.95	1.58
Liaoning	0.8	2.31	3.11	0.62	5.03	1.43
Shandong	0.82	2.27	2.3	0.62	5.11	1.02
Sichuan	0.81	2.15	3.33	0.65	4.89	1.47
Xizang	0.78	3.19	2.98	0.57	5.35	1.78
Yunnan	0.8	2.29	2.87	0.6	5.23	1.26
Zhejiang	0.84	2.11	4.89	0.74	2.85	3.61

To establish a robust comparison, the PTFT method was selected as the primary baseline because it represents the most widely adopted and highly competitive model-based transfer learning paradigm in soil spectroscopy. More importantly, by utilizing the exact same 1D-CNN backbone, the same source-domain pre-training data, and the same target-domain fine-tuning samples, the comparison strictly isolates the effect of the transfer strategy (i.e., global fine-tuning vs. subspace-based class-specific matching) while controlling for network architecture and data volume. This controlled design ensures that the observed performance differences are attributed to the framework itself rather than extraneous factors.

Overall, the class transfer model significantly outperforms the PTFT method across all evaluation metrics on the identical test sets. The average R^2 of the class transfer model across all provinces in China was 0.81, ranging from 0.77 to 0.84. In contrast, the average R^2 of the PTFT method was only 0.66, with a range between 0.57 and 0.74. All provinces achieved an R^2 above 0.77 using the class transfer model, whereas the PTFT method had R^2 values below 0.70 in 10 provinces, indicating that the former possesses stronger cross-domain fitting capability. Regarding RMSE, the class transfer model had an average RMSE of 2.33 g kg^{-1}, with the lowest value of 2.11 g kg^{-1} in Henan and Zhejiang, and the highest of 3.19 g kg^{-1} in Xizang. The PTFT method had a much higher average RMSE of 4.70 g kg^{-1}, with the minimum at 2.85 g kg^{-1} in Zhejiang and the maximum reaching 5.35 g kg^{-1} in Xizang. The average RPD of the class transfer model was 3.24, ranging from 2.30 to 4.80. The PTFT method's average RPD was only 1.71, ranging from 1.02 to 3.61. Therefore, compared to the global transfer model established by PTFT, the class transfer model improved R^2 by 5.48–47.37%, reduced RMSE by 25.96–94.24%, and increased RPD by 35.46–125.49%. Notably,

in the PTFT method, provinces such as Shandong, Hebei, and Liaoning had RPD values below 1.5, indicating poor predictive performance, whereas the class transfer model achieved RPD values greater than 2.3 across all provinces, with nine provinces exceeding 2.5 and six surpassing 3.0. It is important to emphasize that the experimental comparison is not limited to a single scenario. The evaluation spans 11 distinct provinces across China, covering highly heterogeneous pedogenic environments, from the arid regions of Xinjiang to the cold regions of Heilongjiang and the humid regions of Zhejiang. The consistent and substantial outperformance of the class-specific transfer model across all 11 independent provinces (improving R^2 by 5.48–47.37% and RPD by 35.46–125.49%) provides strong empirical evidence of its robustness and generalization capability, demonstrating that the superiority is not an artifact of a specific dataset but a systematic advantage under heterogeneous conditions. The above comparison clearly demonstrates that the transfer framework based on subspace and class transfer models substantially mitigates the distribution differences between the source domain (SD) and target domain (TD) through spectral clustering matching and class-specific model specialization, and its transfer performance significantly outperforms the conventional PTFT method under the tested cross-regional conditions.

4. Discussion

4.1. Subspace-Based Transfer Learning Framework

Over the past two decades, soil hyperspectral technology has made significant progress and has been proven to be an effective method for predicting soil properties [6,12,13,18,21,50,59]. However, due to the spatial heterogeneity of soil, a model developed for a specific area cannot be applied to another area [14,16]. To address this issue, researchers have introduced transfer learning (TL) into the field, aiming to enhance the transferability of inversion models [15,17]. However, despite numerous studies using TL algorithms to build transfer models with certain transfer capabilities, the results have not been satisfactory [15,24]. For this reason, the traditional approach to building soil transfer models has been to develop a single global transfer model, expecting a unified spectral-physicochemical value mapping relationship to apply to all soil samples [60,61]. However, soils exhibit "same spectrum, different properties" and "different spectra, same properties" phenomena. The former phenomenon refers to soils formed under different regions and pedogenic conditions that may have very similar spectral reflectance curves but significantly differ in actual physical and chemical properties; the latter is the opposite—soils with distinctly different spectral characteristics may actually have similar physical and chemical properties [15,17,24,60,61]. Global transfer models attempt to cover all these complex relationships using a fixed one-to-one mapping transfer, which leads to spurious correlations when dealing with "same spectrum, different semantics" and fails to effectively exploit underlying patterns in cases of "different spectrum, same semantics," ultimately causing linear models to fail and severely compromising the model's generalization ability and cross-region transferability. Second, global transfer models may also sacrifice local optimal performance, as the global model attempts to adapt a unified mapping between spectral and physicochemical values across all regions and soil types. To achieve acceptable overall performance, the model inevitably "averages" the characteristics of different subregions. This issue persists even after fine-tuning on the target domain (TD) [20,51]. For example, it might abandon certain highly sensitive pattern rules in a specific region because those rules are either invalid or even disruptive in other regions. Consequently, the model's prediction accuracy on any individual subregion is typically lower than that of a local model trained specifically for that subregion alone [19]. Second, universal models aim to minimize overall error rather than minimizing error in every local region [47]. Therefore, it inevitably sacrifices some of the local optimal fitting capability. Following the proposed approach, autoencoder clustering is employed to

map high-dimensional nonlinear spectra into a low-dimensional feature space, enabling the identification of groups with authentic spectral similarity. Independent modeling is then conducted within each group, thereby implementing a divide-and-conquer strategy. Each cluster sub-model in the Global Soil Spectral Library (GSSL) preserves a certain "typical spectral pattern" on a global scale, while the matching process activates only the most relevant sub-models for Chinese samples. This enables on-demand extraction of locally applicable knowledge from the global knowledge base, effectively balancing generalization capability with specific feature capture.

4.2. The Problem of Negative Transfer in Transfer Learning

In predicting soil properties using soil hyperspectral data, the traditional model-based transfer learning (pre-training and fine-tuning) approach is prone to "negative transfer," primarily due to limitations in how it handles the complexity and inherent heterogeneity of soil spectral data. The transfer strategy of pre-training and fine-tuning (PTFT) essentially assumes that the data from the source domain (SD) and the target domain (TD) share a globally unified feature space [62,63]. However, the formation mechanism of soil spectral data is extremely complex, as its reflectance characteristics are jointly influenced by various physical and chemical properties such as soil moisture content, mineral composition, texture, organic matter content, and surface roughness, with these effects exhibiting highly coupled and nonlinear interactions across different spectral bands [24,64,65]. For example, moisture is one of the key factors determining soil reflectance; as soil water content increases, its spectral curve from visible to near-infrared generally shows a downward trend [64]. In this case, when the SD is derived from a comprehensive and broad general soil spectrum sourced from diverse climatic zones and parent materials worldwide, the spectral information it contains represents a highly mixed, broad statistical commonality designed to cover all possible spectral forms. If such a general model is directly fine-tuned for a specific TD with a particular soil-forming environment, it is highly likely that the model will learn patterns in the deep layers that contradict the characteristics of the TD, as it attempts to simultaneously adapt to the various redundant features of SD and the specificity of TD [66–68]. Meanwhile, the fine-tuning process may over-adjust parameters that are effective for SD but represent "noise" for TD. For example, certain absorption features emphasized to fit high-organic-matter samples in SSD might actually correspond to noise from low-organic-matter, high-iron-oxide samples in TD. This global parameter adjustment causes confusion, so that knowledge from SSD not only fails to assist learning but instead severely disrupts the inherent spectral property relationships within TD, ultimately resulting in degraded prediction performance known as "negative transfer".

In contrast, the TL strategy based on subspace and class transfer models accurately identifies and exploits the inherent "clustering" or "subdomain" structure of soil spectral data. In soil science, this is equivalent to classifying SD data according to their intrinsic spectral patterns, so-called "spectral soil types" shaped collectively by various pedogenic factors such as climate, parent material, topography, and time [25,69,70]. By clustering in the SD, the algorithm can decompose the originally seemingly chaotic global spectral space into several subspaces with similar spectral characteristics, each representing a similar physicochemical environment [19]. Subsequently, an inversion model is established separately for each subspace, effectively enabling each model to focus solely on learning the relationship between spectra and SOM under one or several specific soil-forming conditions. When an unknown sample from TD arrives, the transfer process does not directly apply the global model but first matches it using a clustering model, specifically determining which "spectral soil type" in SSD the target sample's spectral morphology is most similar to. Only samples matched to the corresponding group will trigger the group-specific inver-

sion model for prediction. This mechanism cleverly addresses the core issue of "negative transfer," namely, differences in domain distributions [23,35,49,68]. Since each sub-model is trained on a highly homogeneous subset of data, the feature relationships within the model are extremely pure and stable. When TD samples are assigned to a particular group, their spectral characteristics and attribute relationships closely resemble those of the training data in that group, effectively performing an accurate "find-similar" process and fundamentally avoiding interference from environmental noise across different conditions in global transfer. This strategy is particularly reasonable in soil spectral prediction because two samples from entirely different parent material regions may exhibit vastly different spectral curves, even if their organic matter contents are similar; conversely, within the same region, due to similar pedogenic processes, spectral characteristics tend to be relatively consistent. This aligns with the third law of geography [71]. Thus, clustering-based group modeling essentially transfers "environmental knowledge" that aligns with the third law of geography, rather than "universal knowledge," thereby successfully addressing the challenge of negative transfer caused by global data heterogeneity.

4.3. Class-Based Transfer Model and Single Global Transfer Model

From a model perspective, the superior transfer performance of subspace-based and class-based TL over global pre-training followed by fine-tuning stems fundamentally from the inherent differences in how these two strategies construct hypothesis spaces and the degrees of freedom in parameter updates [72,73]. A global transfer model, such as a deep neural network, is pre-trained by searching for a globally optimal parameter point over the entire data distribution of the source domain (SD), which minimizes the average loss across all samples in the SD [52]. However, when the SD data exhibits a multi-modal distribution (i.e., multiple distinct subgroups), this global optimal solution actually becomes a "compromise solution," where neurons within the model must simultaneously respond to feature patterns from different subgroups. As a result, the extracted feature representations become globally shared representations that possess some discriminative power across all groups but lack sufficient precision for any specific subgroup [52,73]. When fine-tuning the model to adapt to TD, gradient descent updates perform a local search within the neighborhood of this globally optimal parameter point. Since TD data typically correspond to only a single subgroup within SD, the fine-tuning process attempts to forcibly alter the parameters of neurons originally used to respond to other unrelated subgroups, in order to accommodate the characteristics of the current group [71,74]. The problem stems from the fact that pre-trained parameters have already locked the entire network into a global loss sensitive basin. Drastic parameter adjustments quickly disrupt the knowledge learned from the source domain, yet modest adjustments fail to suppress noise interference from other subgroups. As a result, the model ends up in a suboptimal equilibrium point that is neither globally nor locally optimal, thereby causing negative transfer.

In contrast, the class-specific transfer model fundamentally decouples the hypothesis space across different subgroups' transfer. It clusters TD data into several compact convex subsets and independently trains a local model for each subset (which can be viewed as a simple base learner on a low-dimensional manifold) [19]. From the perspective of function approximation, each local model is constrained within the closed convex hull of its corresponding subgroup, so its parameters do not need to carry redundant information across groups. As a result, the complexity of the hypothesis space and the upper bound on generalization error are significantly lower than those of a global model [75]. Upon arrival of a target domain sample, the clustering matching mechanism implements a rigid model selection process. It matches the target sample's input space position against the applicable domains of individual local models via nearest neighbor comparison, activating solely

the corresponding local model. This approach precludes cross-group parameter coupling, whereby a single weight in a global model would otherwise need to simultaneously accommodate multiple groups. Consequently, each parameter within a local model responds exclusively to variations indigenous to its own group [6,23,42,76]. More importantly, since the training of local models does not require considering data from other groups, their loss functions are convex or approximately convex, enabling convergence to the global optimum within that subspace [36,38]. Therefore, for samples in the test data (TD), as long as they can be accurately matched to a group in the source data (SD) with a similar conditional distribution, predicting using that group's local model results in an expected error approximately equal to the intrinsic noise level within the group, without introducing systematic bias from other groups [77,78]. This achieves an optimal balance between model complexity and data alignment, enabling the transfer performance of class-specific transfer models to significantly outperform the fine-tuning paradigm transfer, which is constrained by global coupling and parameter interference.

To substantiate the claimed mechanism without introducing additional external baselines, the contribution of the two core components (i.e., the AEC clustering and the class-specific local modeling) can be systematically inferred from the existing experimental results. First, regarding the clustering component: the PCA distribution (Figure 4) and the intra-cluster SOM statistics (Table 2) demonstrate that AEC effectively partitions the heterogeneous GSSL data into 18 distinct and internally homogeneous subdomains. If this clustering component were absent, the model would degenerate into a global transfer model. As evidenced by the PTFT baseline in Table 3, such a global model suffers from severe performance degradation due to parameter coupling and the "averaging" effect. Second, regarding the local modeling component: if the clustering was retained but the models were not decoupled (i.e., if a single global model was forced to fit all clusters), the hypothesis space complexity would remain unchanged. The high prediction accuracy of the 18 independent models within their respective clusters (Figures 5 and 8) confirms that spatial decoupling and independent parameter optimization are what actually drive the performance gains. Therefore, the synergy between AEC clustering and class-specific modeling—rather than either component in isolation—is the primary mechanism.

While recent domain adaptation techniques (e.g., CORAL, MMD, or adversarial training) have been proposed, they often require complex modifications to the network architecture and may struggle with highly nonlinear spectral data where the marginal distributions of SD and TD differ drastically. In contrast, the proposed framework provides a more interpretable and computationally feasible solution by directly partitioning the data space. Although the present comparison focuses on a single PTFT baseline, this baseline represents the strongest and most natural standard in soil spectral transfer. The highly consistent performance gains across three evaluation metrics and 11 independent regions strongly substantiate the effectiveness of the framework. Furthermore, we have explicitly moderated the claim of "superiority" to "significant advantages under the tested cross-regional settings," acknowledging that broader comparisons remain an important avenue for future work.

4.4. Limitations and Future Work

This study proposes a soil spectral transfer learning framework based on subspace and class transfer models, grounded in clustering algorithms, soil spectral feature similarity, and the third law of geography, and it validates its feasibility and transfer potential using the Global Soil Spectral Library (GSSL) and the Chinese Soil Spectral Library (CSSL). The research findings provide new insights for the practical application of soil hyperspectral data, digital soil mapping, and the local utilization of large spectral libraries. However, limitations still exist. First, the hard assignment in clustering algorithms introduces prediction risks for

"boundary samples". Target domain (TD) samples are forced to be assigned to the nearest cluster, but if they lie in a vague region between two clusters or are far from all cluster centers, predicting them with a single sub-model can easily lead to significant errors [19,25,77]. Second, if a certain category has very few samples in the global database (such as some cold or extreme soil types), the corresponding sub-model will be severely overfitted, making predictions for Chinese samples belonging to that category unreliable [19,25]. Third, the successful assignment of all CSSL samples to GSSL-derived clusters should not be interpreted as evidence of cross-database generalization. Because each target sample is forced to its nearest cluster, high assignment coverage mainly reflects the inclusiveness of the cluster centers, and it does not guarantee that the matched class transfer model will produce accurate predictions [25]. The actual transfer performance is therefore evaluated separately through the prediction results in Section 3.4, and the distinction between assignment coverage and generalization performance should be kept explicit in future applications [27]. Fourth, although the proposed framework was extensively validated across 11 provinces against a strong PTFT baseline, the experimental comparison did not include other advanced domain adaptation methods (e.g., CORAL, MMD, or DANN) or multi-source domain generalization frameworks. Future research should incorporate these additional baselines to provide more comprehensive benchmarking, which will further consolidate the theoretical and practical value of the proposed approach. Finally, and importantly, the issue of dataset harmonization must be acknowledged as a limitation of this study. The GSSL and CSSL datasets are compiled from multiple sources worldwide and across China, meaning the spectra were obtained under different measurement conditions and with different spectrometers [27,28]. Similarly, the reference SOM values were measured using different laboratory methods. Although the proposed AEC clustering and class-specific modeling framework demonstrated strong transferability across these heterogeneous datasets (which to some extent reflects the robustness of the data-driven approach), the lack of rigorous spectral harmonization (e.g., instrument response matching, spectral resampling) and chemical protocol alignment introduces uncertainties in the calibration and validation process. Future research should prioritize cross-instrument harmonization and incorporate measurement uncertainty into the transfer learning framework to further improve its scientific rigor.

The future should move from "hard matching" to "soft weighted integration," avoiding forced classification by calculating the degree of membership of a target sample to each cluster center, and using the weighted average of outputs from N sub-models as the final prediction. This helps smooth boundary samples and improve stability. Additionally, weak constraints based on physicochemical data can be introduced during autoencoder (AEC) clustering—for example, ensuring that soils within the same cluster tend to have similar parent materials or diagnostic horizons, rather than relying solely on spectral similarity [21,28,78]. A multi-task autoencoder can be considered to simultaneously reconstruct spectra and predict approximate physicochemical ranges, thereby obtaining more transferable features.

5. Conclusions

To resolve the key challenges inherent in cross-regional soil hyperspectral modeling, namely, the inability of global transfer models to adequately capture nonlinear spectral attribute relationships and their vulnerability to negative transfer resulting from substantial distributional disparities between the source domain and target domain, this paper presents a transfer learning modeling framework founded on subspace and class transfer models. The framework clusters SD samples based on spectral similarity, independently constructs class-specific transfer models for each group, and achieves cross-database prediction by matching TD samples to corresponding groups. The framework was validated using the Global Soil Spectral Library (GSSL) as the source data (SD), the Chinese Soil Spectral Library (CSSL)

as the target data (TD), and soil organic matter (SOM) as the target attribute, leading to the following main conclusions. (1) Autoencoder (AEC) unsupervised clustering enables high-quality partitioning in GSSL and assigns all 3735 CSSL samples to GSSL-derived clusters without rejection. This high assignment coverage indicates that the GSSL spectral subspaces are inclusive of the CSSL samples, but it should not be equated with cross-database generalization performance. The spectral features of the SD and TD samples show high overlap within shared groups, establishing a reliable grouping foundation for class transfer model transfer. (2) The proposed subspace and class transfer model transfer framework significantly outperforms traditional model-based pre-training and fine-tuning (PTFT) global transfer models. In CSSL, the average R^2 improvement of the class transfer model ranges from 5.48% to 47.37%, *RMSE* reduction ranges from 25.96% to 94.24%, and *RPD* improvement ranges from 35.46% to 125.49%. (3) This framework effectively addresses the negative transfer problem caused by the mismatch between SD and TD distributions, providing strong empirical support for the assumption that "class-specific transfer models" can outperform "global transfer models" under the tested heterogeneous cross-regional conditions. This study provides a novel methodological framework for cross-regional soil hyperspectral detection, expanding the application boundaries of transfer learning and large-scale spectral databases in heterogeneous geological settings, and shows promise in advancing digital soil mapping from a global model to a class-based model paradigm.

Author Contributions: Conceptualization, J.D., J.W., H.L. and D.L.; methodology, J.D., J.W. and H.L.; software, B.H.; validation, J.L., D.L., J.P., and Z.S.; formal analysis, J.D., J.W., H.L., J.L., D.L., J.P. and Z.S.; investigation, J.D., J.W., H.L., J.L., D.L., J.P. and Z.S.; resources, B.H.; data curation, J.D., J.W. and H.L.; writing—original draft preparation, J.D., J.W. and H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by grants from the Bingtuan Science and Technology Program (Grant Nos. 2025DA002), the National Science Foundation of China (Grant Nos. 42261016), and the Jiangxi "Double Thousand plan" (Nos. jxsq202301091).

Data Availability Statement: The authors do not have permission to share the data.

Acknowledgments: The authors appreciate all the data provided by each open database. The authors thank the anonymous reviewers and academic editors for their comments.

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

Abbreviations

The following abbreviations are used in this manuscript:

SD	source domain
TD	target domain
GSSL	the Global Soil Spectral Library
CSSL	the Chinese Soil Spectral Library
SOM	soil organic matter
ML	machine learning
DL	deep learning
TL	transfer learning
ISRIC	International Soil Reference and Information Centre
Vis-NIR	visible—near-infrared
AEC	adaptive embedding clustering
PTFT	pre-training and fine-tuning
K-S	Kennard–Stone
FC	multiple fully connected layers
MSE	mean squared error

(KL divergence)	Kullback–Leibler divergence
GA	genetic algorithm
PCA	principal component analysis
CNN	convolutional neural network
SG	Savitzky–Golay
SNV	standard normal variate

References

1. Lal, R. Soil carbon sequestration impacts on global climate change and food security. *Science* **2004**, *304*, 1623–1627. [CrossRef]
2. Bouma, J. Soil security in sustainable development. *Soil Syst.* **2019**, *3*, 5. [CrossRef]
3. Hong, Y.S.; Chen, S.C.; Hu, B.F.; Wang, N.; Xue, J.; Zhuo, Z.Q.; Yang, Y.Y.; Chen, Y.Y.; Peng, J.; Liu, Y.L.; et al. Spectral fusion modeling for soil organic carbon by a parallel input-convolutional neural network. *Geoderma* **2023**, *437*, 0016–7061. [CrossRef]
4. Biney, J.K.M.; Blöcher, J.R.; Borůvka, L.; Vašát, R. Does the limited use of orthogonal signal correction pre-treatment approach to improve the prediction accuracy of soil organic TCArbon need attention? *Geoderma* **2021**, *388*, 114945. [CrossRef]
5. Chen, S.C.; Xu, H.; Xu, D.H.; Ji, W.J.; Li, S.; Yang, M.H.; Shi, Z. Evaluating validation strategies on the performance of soil property prediction from regional to continental spectral data. *Geoderma* **2021**, *400*, 115159. [CrossRef]
6. Liu, J.J.; Hong, Y.S.; Hu, B.F.; Chen, S.C.; Deng, J.; Yin, K.Y.; Lin, J.; Luo, D.F.; Peng, J.; Shi, Z. Hyperspectral inversion of soil organic matter based on improved ensemble learning method. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2025**, *339*, 1386–1425. [CrossRef]
7. Griffiths, P. *Introduction to the Theory and Instrumentation for Vibrational Spectroscopy*; John Wiley & Sons, Ltd.: Chichester, UK, 2010. [CrossRef]
8. Soriano-Disla, J.M.; Janik, L.J.; Viscarra Rossel, R.A.; Macdonald, L.M.; McLaughlin, M.J. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Appl. Spectrosc. Rev.* **2014**, *49*, 139–186. [CrossRef]
9. Hu, B.F.; Chen, S.C.; Hu, J.F.; Xu, J.T.; Li, Y.; Shi, Z. AppliTCAtion of portable XRF and VNIR sensors for rapid assessment of soil heavy metal pollution. *PLoS ONE* **2017**, *12*, e0172438. [CrossRef]
10. Zhong, C.; Li, G.; Meng, Z. Beluga whale optimization: A novel nature-inspired metaheuristic algorithm. *Knowl.-Based Syst.* **2022**, *251*, 109215. [CrossRef]
11. Ahmadi, A.; Emami, M.; Dactcache, A.; He, L. Soil properties prediction for precision agriculture using visible and near-infrared spectroscopy: A systematic review and meta analysis. *Agronomy* **2021**, *11*, 433. [CrossRef]
12. LeCun, Y.; Bengio, Y.; Hinton, G.E. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
13. Singh, S.; Kasana, S.S. Estimation of soil properties from the EU spectral library using long short-term memory networks. *Geoderma Reg.* **2019**, *18*, e00233. [CrossRef]
14. Tsakiridis, N.; Keramaris, K.D.; Theocharis, J.B.; Zalidis, G.C. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. *Geoderma* **2020**, *367*, 114208. [CrossRef]
15. Rossel, R.A.V.; Behrens, T.; Ben-Dor, E.; Chabrillat, S.; Demattê, J.A.M.; 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]
16. Biney, J.K.M.; Vašát, R.; Blöcher, J.R.; Borůvka, L.; Němeček, K. Using an ensemble model coupled with portable X-ray fluorescence and visible near-infrared spectroscopy to explore the viability of mapping and estimating arsenic in an agricultural soil. *Sci. Total Environ.* **2022**, *818*, 151805. [CrossRef]
17. Rossel, R.A.V.; Shen, Z.; Lopez, L.R.; Behrens, T.; Shi, Z.; Wetterlind, J.; Sudduth, K.A.; Stenberg, B.; Guerrero, C.; Gholizadeh, A.; et al. An imperative for soil spectroscopic modelling is to think global but fit local with transfer learning. *Earth-Sci. Rev.* **2024**, *254*, 104797. [CrossRef]
18. Ke, Z.Y.; Shilin, R.; Liang, Y. Advancing soil property prediction with encoder-decoder structures integrating traditional deep learning methods in Vis-NIR spectroscopy. *Geoderma* **2024**, *449*, 117006. [CrossRef]
19. Deng, J.; Hong, Y.S.; Hu, B.F.; Chen, S.C.; Liu, J.J.; Wang, Y.; Peng, J.; Shi, Z. Improving hyperspectral prediction accuracy of soil organic matter using Novel cascaded optimization methods and super ensemble model. *Comput. Electron. Agric.* **2026**, *249*, 111841. [CrossRef]
20. Wang, Y.; Yin, K.Y.; Hu, B.F.; Hong, Y.S.; Chen, S.C.; Liu, J.; Yang, L.L.; Peng, J.; Shi, Z. Ensemble and transfer learning of soil inorganic TCArbon with visible near-infrared spectra. *Geoderma* **2025**, *456*, 117257. [CrossRef]
21. Hong, Y.S.; Jonathan, S.; Tomislav, H.; Chen, S.C.; Wang, N.; Xue, J.; Zhuo, Z.Q.; Peng, J.; Chen, Y.Y.; Liu, Y.; et al. Potential of globally distributed topsoil mid-infrared spectral library for organic carbon estimation. *CATENA* **2024**, *235*, 107628. [CrossRef]
22. Ramirez-Lopez, L.; Wadoux, A.M.J.C.; Franceschini, M.H.D.; Terra, F.S.; Marques, K.P.P.; Sayão, V.M.; Demattê, J.A.M. Robust soil mapping at the farm sTCAle with vis- NIR spectroscopy. *Eur. J. Soil Sci.* **2019**, *70*, 378–393. [CrossRef]

23. Bai, Z.J.; Chen, S.C.; Hong, Y.S.; Hu, B.F.; Luo, D.F.; Peng, J.; Shi, Z. Estimation of soil inorganic TCArbon with visible near-infrared spectroscopy coupling of variable selection and deep learning in arid region of China. *Geoderma* **2023**, *437*, 116589. [CrossRef]
24. Xia, K.; Wu, T.X.; Zhang, S.W.; Wang, S.D.; Li, X.T.; Shen, Q.; Li, T. A new method for high-precision estimation of soil organic matter using two-dimensional correlation spectroscopy-to support collaborative use of global open soil spectral library. *Geoderma* **2024**, *445*, 116877. [CrossRef]
25. Shi, Z.; Wang, Q.; Peng, J.; Ji, W.; Liu, H.; Li, X.; Viscarra Rossel, R.A. Development of a national VNIR soil-spectral library for soil classification and prediction of organic mat ter concentrations. *Sci. China Earth Sci.* **2014**, *57*, 1671–1680. [CrossRef]
26. Sila, A.M.; Shepherd, K.D.; Pokhariyal, G.P. Evaluating the utility of mid-infrared spectral subspaces for predicting soil properties. *Chem. Intell. Lab. Syst.* **2016**, *15*, 92–105. [CrossRef]
27. Hong, Y.; Yu, L.; Chen, Y.; Liu, Y.; Liu, Y.; Cheng, H. Prediction of soil organic matter by VIS–NIR spectroscopy using normalized soil moisture index as a proxy of soil moisture. *Remote Sens.* **2018**, *10*, e28. [CrossRef]
28. Ward, K.J.; Chabrillat, S.; Neumann, C.; Foerster, S. A remote sensing adapted approach for soil organic TCArbon prediction based on the spectrally clustered LUTCAS soil database. *Geoderma* **2019**, *353*, 297–307. [CrossRef]
29. Demattê, J.A.; Dotto, A.C.; Paiva, A.F.; Sato, M.V.; Dalmolin, R.S.; de Araújo, M.D.S.B.; da Silva, E.B.; Nanni, M.R.; Caten, A.T.; Noronha, N.C.; et al. The Brazilian soil spectral library (BSSL): A general view, appliTCAtion and challenges. *Geoderma* **2019**, *354*, e113793. [CrossRef]
30. Breure, T.S.; Prout, J.M.; Haefele, S.M.; Milne, A.E.; Hannam, J.A.; Moreno-Rojas, S.; Corstanje, R. Comparing the effect of different sample conditions and spectral libraries on the prediction accuracy of soil properties from near- and mid-infrared spectra at the field-scale. *Soil Tillage Res.* **2022**, *215*, 105196. [CrossRef]
31. Qiao, H.; Shi, X.; Chen, H.; Lyu, J.; Hong, S. Effective prediction of soil organic matter by deep SVD concatenation using FT-NIR spectroscopy. *Soil Tillage Res.* **2022**, *215*, 105223. [CrossRef]
32. Clingensmith, C.M.; Grunwald, S. Predicting soil properties and interpreting vis-NIR models from across Continental United States. *Sensors* **2022**, *22*, 3187. [CrossRef]
33. Dotto, A.C.; Dalmolin, R.S.D.; Grunwald, S.; ten TCAten, A.; Pereira Filho, W. Two pre processing techniques to reduce model covariables in soil property predictions by Vis-NIR spectroscopy. *Soil Tillage Res.* **2017**, *172*, 59–68. [CrossRef]
34. Seema, G.A.K.; Das, B.S.; Reddy, N. AppliTCAtion of VIS-NIR spectroscopy for estimation of soil organic TCArbon using different spectral preprocessing techniques and multivariate methods in the middle Indo-Gangetic plains of India. *Geoderma* **2020**, *23*, e00349. [CrossRef]
35. Wang, Y.; Chen, S.C.; Hong, Y.S.; Hu, B.F.; Peng, J.; Shi, Z. A comparison of multiple deep learning methods for predicting soil organic TCArbon in Southern Xinjiang, China. *Comput. Electron. Agric.* **2023**, *212*, 108067. [CrossRef]
36. Huang, P.H.; Huang, Y.; Wang, W.; Wang, L. Deep embedding network for clustering. In *22nd International Conference on Pattern Recognition*; IEEE: New York, NY, USA, 2014; pp. 1532–1537. [CrossRef]
37. Ghasedi, D.A.K.; Herandi, C.; Deng, W.; TCAi, H. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2017; pp. 5736–5745. [CrossRef]
38. Deng, J.; Liu, J.J.; Li, H.X.; Hu, B.F.; Peng, J. Spectral Similarity-Based Grouping Improves Laboratory Vis-NIR Prediction of Soil Organic Matter Under Soil Heterogeneity. *Sensors* **2026**, *26*, 6824. [CrossRef]
39. Esmaeili, M.; Abbasi-Moghadam, D.; Sharifi, A.; Tariq, A.; Li, Q. Hyperspectral image band selection based on CNN embedded GA (CNNeGA). *J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1927–1950. [CrossRef]
40. Yin, K.Y.; Deng, J.; Li, H.X.; Feng, C.H.; Peng, J. A Novel Self-Attention Mechanism-Based Dynamic Ensemble Model for Soil Hyperspectral Prediction. *Sensors* **2026**, *26*, 195. [CrossRef]
41. Ben-Dor, E. Quantitative remote sensing of soil properties. *Adv. Agron.* **2002**, *75*, 173–243. [CrossRef]
42. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1990. [CrossRef]
43. Maćkiewicz, A.; Ratajczak, W. Principal components analysis (PCA). *Comput. Geosci.* **1993**, *19*, 303–342. [CrossRef]
44. Berger, V.W.; Zhou, Y. Kolmogorov–Smirnov Test: Overview. In *Wiley Statsref: Statistics Reference Online*; Wiley: Hoboken, NJ, USA, 2014. [CrossRef]
45. Lever, J.; Krzywinski, M.; Altman, N. Principal component analysis. *Nat. Methods* **2017**, *14*, 641–642. [CrossRef]
46. Boško, M.; Aleksandra, B. Organic TCArbon estimation in a regional soil Vis-NIR database supported by unsupervised learning and chemometrics techniques. *Soil Adv.* **2024**, *2*, 100013. [CrossRef]
47. Yang, J.; Wang, X.; Wang, R.; Wang, H. Combination of convolutional neural networks and recurrent neural networks for predicting soil properties using Vis–NIR spectroscopy. *Geoderma* **2020**, *380*, 114616. [CrossRef]
48. Salehi-Varnousfaderani, B.; Honarbakhsh, A.; Tahmoures, M.; Akbari, M. Soil erodibility prediction by Vis-NIR spectra and environmental covariates coupled with GIS, regression and PLSR in a watershed scale, Iran. *Geoderma Reg.* **2022**, *28*, e00470. [CrossRef]

49. Singha, C.; Swain, K.C.; Sahoo, S.; Govind, A. Prediction of soil nutrients through PLSR and SVMR models by VIs-NIR reflectance spectroscopy. *Egypt. J. Remote Sens. Space Sci.* **2023**, *26*, 901–918. [CrossRef]
50. Wang, Z.; Chen, S.C.; Lu, R.; Zhang, X.L.; Ma, Y.X.; Shi, Z. Non-linear memory-based learning for predicting soil properties using a regional vis-NIR spectral library. *Geoderma* **2024**, *441*, 116752. [CrossRef]
51. Bu, Y.; Luo, J.; Tian, Q.; Li, J.; Cao, M.; Yang, S.; Guo, W. Nondestructive detection of internal quality in multiple peach varieties by Vis/NIR spectroscopy with multi-task CNN method. *Postharvest Biol. Technol.* **2025**, *227*, 113579. [CrossRef]
52. Hosna, A.; Merry, E.; Gyalmo, J.; Alom, Z.; Aung, Z.; Azim, M.A. Transfer learning: A friendly introduction. *J. Big Data* **2022**, *9*, 102. [CrossRef]
53. You, K.; Kou, Z.; Long, M.; Wang, J. Co-tuning for transfer learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17236–17246.
54. Zeng, R.; Zhang, J.P.; TCAi, K.; Gao, W.C.; Pan, W.J.; Jiang, C.Y.; Zhang, P.Y.; Wu, B.W.; Wang, C.H.; Jin, X.Y.; et al. How similar is "similar", or what is the best measure of soil spectral and physiochemiTCAl similarity? *PLoS ONE* **2021**, *16*, e0247028. [CrossRef]
55. Rousseeuw, P.J. Silhouettes: A graphiTCAl aid to the interpretation and validation of cluster analysis. *J. Comp. Appl. Maths* **1987**, *20*, 53–65. [CrossRef]
56. Su, W.H.; Jessada, K.; Youji, K.; Chawalit, J. A hybrid\K-means and KNN approach for enhanced short-term load foreTCAsting incorporating holiday effects. *Energy Rep.* **2024**, *12*, 5942–5959. [CrossRef]
57. Delimiro, V.-T.; Enrique, D.-D.; Adel, M.-M. A multiple factor analysis and hierarchiTCAl clustering of global logistics governance and development. *Decis. Anal. J.* **2025**, *15*, 100579. [CrossRef]
58. Kvålseth, T.O. TCAutionary note about R^2. *AmeriTCAn Stat.* **1985**, *39*, 279–285. [CrossRef]
59. Hamed, T.; José, C.; Marmar, S.; Sebastian, V. Predicting key soil properties from Vis-NIR spectra by applying dual-wavelength indices transformations and stacking machine learning approaches. *Soil Tillage Res.* **2023**, *229*, 105684. [CrossRef]
60. Padarian, J.; Minasny, B.; McBratney, A. Transfer learning Sto localise a continental soil Vis-NIR calibration model. *Geoderma* **2019**, *340*, 279–288. [CrossRef]
61. Shen, Z.; Ramirez-Lopez, L.; Behrens, T.; Cui, L.; Zhang, M.; Walden, L.; Wetterlind, J.; Shi, Z.; Sudduth, K.A.; Baumann, P.; et al. Deep transfer learning of global spectra for local soil carbon monitoring. *ISPRS J. Photogramm. Remote Sens.* **2022**, *188*, 190–200. [CrossRef]
62. Torrey, L.; Shavlik, J. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*; IGI Global Scientific Publishing: Hershey, PA, USA, 2010; pp. 242–264. [CrossRef]
63. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 9. [CrossRef]
64. Agnese, B.; Luca, B.; Mario, M.; Felice, D.O. From pre-training to fine-tuning: An in-depth analysis of Large Language Models in the biomedical domain. *Artif. Intell. Med.* **2024**, *157*, 103003. [CrossRef]
65. Luca, R.; Harry, K.; Feng, Y.S.; Jing, K.J.; Zhang, D.D. Accelerating bioprocess digital twin development by integrating hybrid modelling with transfer learning. *Chem. Eng. J.* **2025**, *511*, 162018. [CrossRef]
66. Hajek, P.; Hikkerova, L.; Sahut, J.M. Fake review detection in e-Commerce platforms using aspect-based sentiment analysis. *J. Bus. Res.* **2023**, *167*, 114143. [CrossRef]
67. Jin, W.; Zhao, B.; Zhang, L.; Liu, C.; Yu, H. Back to common sense: Oxford dictionary descriptive knowledge augmentation for aspect-based sentiment analysis. *Inf. Process. Manag.* **2023**, *60*, 103260. [CrossRef]
68. Gaurav, D.; Anil, K.D.; Kamaljit, K.; Gaurav, R.; Parveen, K. Adaptive contextual memory graph transformer with domain-adaptive knowledge graph for aspect-based sentiment analysis. *Expert Syst. Appl.* **2025**, *278*, 127300. [CrossRef]
69. Mahmood, H.S.; Hoogmoed, W.B.; Van Henten, E.J. Proximal gamma-ray spectroscopy to predict soil properties using windows and full-spectrum analysis methods. *Sensors* **2013**, *13*, 16263–16280. [CrossRef]
70. Mahmood, H.S.; Hoogmoed, W.B.; Van Henten, E.J.; Adamchuck, V.I. Estimating soil properties with a proximal gamma-ray spectrometer using windows and full-spectrum analysis methods. In Proceedings of the Second Global Workshop on Proximal Soil Sensing, Montreal, QC, Canada, 15–18 May 2011; pp. 132–135.
71. Li, L.Y.; Zhang, Z.P.; Sun, M.L.; Ding, J.L.; Wang, J.Z.; Xu, D.; Huang, Y.Y. Selecting the right samples rather than more samples: A new spectral–environmental similarity strategy for local soil spectral modeling. *Geoderma* **2026**, *467*, 117710. [CrossRef]
72. Patricia, N.; Caputo, B. Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2014; pp. 1442–1449. [CrossRef]
73. Ghafoorian, M.; Mehrtash, A.; Kapur, T.; Karssemeijer, N.; Marchiori, E.; Pesteie, M.; Wells, W.M. Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, 11–13 September 2017, Proceedings, Part III 20*; Springer: Cham, Switzerland, 2017; pp. 516–524. [CrossRef]
74. Akuzawa, K.; Iwasawa, Y.; Matsuo, Y. Adversarial invariant feature learning with accuracy constraint for domain generalization. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 16–20*; Springer: Heidelberg, Germany, 2019; Proceedings, Part II; pp. 315–331. [CrossRef]

75. Affeldt, S.; Labiod, L.; Nadif, M. Spectral clustering via ensemble deep autoencoder learning (SC-EDAE). *Pattern Recognit.* **2020**, *108*, 107522. [CrossRef]
76. Cohen, S. Maximum difference scaling: Improved measures of importance and preference for segmentation. *Sawtooth Softw. Conf. Proc.* **2003**, *530*, 61–74.
77. Yang, X.; Deng, C.; Zheng, F.; Yan, J.; Liu, W. Deep spectral clustering using dual autoencoder network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2019; pp. 4066–4075. [CrossRef]
78. Tavakoli, N.; Siami-Namini, S.; Adl Khanghah, M.; Mirza Soltani, F.; Siami Namin, A. An autoencoder-based deep learning approach for clustering time series data. *SN Appl. Sci.* **2020**, *2*, 937. [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.