Transformer-based Models for High-Resolution Soil Property Mapping: Leveraging Deep Learning and Multi-Modal Remote Sensing for Precision Agriculture

Dr. Anastasios Georgiou [1]*, Dr. Seyed Mahdi Hosseini [2], Dr. Eshetu Tadesse [3]
[1] School of Agricultural Sciences, Aristotle University of Thessaloniki, Greece
[2] Department of Soil Science, Aristotle University of Thessaloniki, Greece
[3] Department of Natural Resource Management, Aristotle University of Thessaloniki, Greece

* Corresponding Author: **Dr. Anastasios Georgiou**

Article Info

P-ISSN: 3051-3448
E-ISSN: 3051-3456
Volume: 06
Issue: 01
January - June 2025
Received: 02-12-2024
Accepted: 02-01-2025
Published: 03-02-2025
Page No: 01-06

Abstract
High-resolution soil property mapping is essential for precision agriculture and sustainable land management, but traditional approaches face limitations in capturing complex spatial relationships and integrating multi-modal data sources. This study presents a novel application of transformer-based deep learning models for predicting soil properties at unprecedented spatial resolution using multi-modal remote sensing data. We developed and evaluated three transformer architectures: Vision Transformer (ViT), Swin Transformer, and a custom Multi-Modal Transformer (MMT) for mapping soil organic carbon (SOC), pH, clay content, and available nitrogen across 15,000 km² of agricultural landscapes in the Midwest USA. The models integrated Sentinel-2 multispectral imagery, Landsat-8 thermal data, ALOS PALSAR-2 synthetic aperture radar, digital elevation models, and 8,247 ground truth soil samples collected from multiple depths (0-15, 15-30, 30-60 cm). Data preprocessing involved advanced augmentation techniques, spatial-temporal feature extraction, and attention-based fusion mechanisms. The Multi-Modal Transformer achieved superior performance with R^2 values of 0.89 for SOC, 0.84 for pH, 0.81 for clay content, and 0.76 for available nitrogen, outperforming traditional machine learning methods (Random Forest: R^2 = 0.72-0.78) and convolutional neural networks (CNN: R^2 = 0.75-0.82). Root mean square errors were reduced by 23-31% compared to conventional approaches. The transformer models demonstrated exceptional capability in capturing long-range spatial dependencies and complex non-linear relationships between soil properties and environmental covariates. Attention mechanism analysis revealed that the models effectively learned to focus on relevant spectral bands, topographic features, and spatial contexts. High-resolution maps (10-meter pixel size) were generated showing detailed spatial variability previously undetectable with traditional methods. Computational efficiency analysis showed 2.3× faster inference compared to equivalent CNN architectures while maintaining superior accuracy. Cross-validation experiments across different agro-ecological zones confirmed model robustness and transferability. The study demonstrates the transformative potential of transformer architectures for digital soil mapping, enabling precision agriculture applications and supporting data-driven decision-making for sustainable soil management.

Keywords: Transformer Models, Soil Property Mapping, Deep Learning, Remote Sensing, Precision Agriculture, Vision Transformer, Attention Mechanism, Digital Soil Mapping, Multi-Modal Data Fusion

Introduction
Accurate mapping of soil properties at high spatial resolution represents a fundamental requirement for precision agriculture, environmental monitoring, and sustainable land management [1].

Traditional soil mapping approaches rely on limited point observations and interpolation techniques that often fail to capture the complex spatial heterogeneity inherent in soil systems [2]. The advent of remote sensing technologies and digital soil mapping has revolutionized soil property prediction, but conventional machine learning methods still face challenges in integrating multi-modal data sources and capturing complex spatial relationships [3].

Deep learning techniques have emerged as powerful tools for soil property mapping, demonstrating superior performance compared to traditional statistical and machine learning approaches [4]. Convolutional Neural Networks (CNNs) have shown particular promise in extracting spatial features from remote sensing imagery and digital elevation models [5]. However, CNNs are limited by their local receptive fields and may struggle to capture long-range spatial dependencies that are crucial for understanding soil formation processes and landscape-scale patterns [6].

Transformer architectures, originally developed for natural language processing, have recently gained attention in computer vision applications due to their ability to model long-range dependencies through self-attention mechanisms [7]. The Vision Transformer (ViT) introduced the concept of treating images as sequences of patches, enabling the application of transformer architectures to image analysis tasks [8]. Subsequent developments including the Swin Transformer have addressed scalability issues and improved performance on various computer vision benchmarks [9].

The self-attention mechanism in transformer models provides several advantages for soil property mapping applications. Unlike CNNs that process information through localized convolutions, transformers can directly model relationships between distant spatial locations, potentially capturing landscape-scale processes that influence soil development [10]. The attention weights provide interpretability by highlighting which spatial regions and features contribute most to predictions, addressing the "black box" nature of many deep learning approaches [11].

Multi-modal data integration represents another critical challenge in digital soil mapping, as soil properties are influenced by diverse environmental factors including climate, topography, geology, vegetation, and land use history [12]. Traditional approaches often struggle to effectively combine information from different sensor types and spatial-temporal scales. Transformer architectures offer promising solutions through their flexible attention mechanisms that can learn optimal weighting schemes for different data modalities [13].

Recent advances in remote sensing technology have provided unprecedented opportunities for soil monitoring through multi-spectral, hyperspectral, thermal, and radar imagery [14]. Sentinel-2 and Landsat-8 missions provide regular global coverage with moderate spatial resolution, while synthetic aperture radar (SAR) data offers weather-independent observations of soil surface conditions [15]. The integration of these diverse data sources requires sophisticated modeling approaches capable of extracting complementary information and handling varying spatial-temporal resolutions [16].

The scalability and computational efficiency of soil mapping approaches are critical considerations for operational applications covering large geographic areas [17]. Traditional machine learning methods may require extensive feature engineering and struggle with high-dimensional data, while deep learning approaches often demand significant computational resources. Transformer models offer potential advantages through their parallel processing capabilities and efficient attention mechanisms [18].

This study aims to evaluate the effectiveness of transformer-based architectures for high-resolution soil property mapping using multi-modal remote sensing data. Specific objectives include: (1) developing and optimizing transformer models for soil property prediction, (2) comparing transformer performance against conventional machine learning and CNN approaches, (3) analyzing attention mechanisms to understand model decision-making processes, (4) generating high-resolution soil property maps for precision agriculture applications, and (5) evaluating computational efficiency and scalability for operational deployment.

Materials and Methods

Study Area and Sampling Design

The research was conducted across agricultural landscapes in the Midwest USA, encompassing portions of Iowa, Illinois, Indiana, and Ohio (39°45'N to 42°30'N, 88°30'W to 91°15'W). The study area covers approximately 15,000 km² and represents diverse soil types including Mollisols, Alfisols, and Entisols formed under varying topographic and climatic conditions. The region experiences continental climate with mean annual precipitation ranging from 800-1,200 mm and temperatures from -5°C to 25°C.

Soil sampling employed stratified random design based on soil survey units, topographic position, and land use categories. A total of 8,247 sampling points were established with minimum spacing of 500 meters to ensure spatial independence. Samples were collected from three depth intervals: 0-15 cm (surface), 15-30 cm (subsurface), and 30-60 cm (subsoil) to capture vertical soil profile variations.

Geographic coordinates were recorded using differential GPS with sub-meter accuracy. Additional site information including land use, crop type, management practices, and surface conditions were documented for each sampling location.

Laboratory Analysis

Soil samples were processed following standardized protocols for four target properties: soil organic carbon (SOC), pH, clay content, and available nitrogen. SOC was determined using dry combustion method with elemental analyzer (Costech ECS 4010, Valencia, CA). Soil pH was measured in 1:1 soil-water suspension using calibrated electrodes. Clay content was quantified through particle size analysis using laser diffraction (Beckman Coulter LS13320, Brea, CA). Available nitrogen was assessed through alkaline permanganate oxidation with colorimetric detection.

All analyses were performed in triplicate with quality control samples comprising 10% of total samples. Inter-laboratory comparison exercises ensured measurement consistency and accuracy across different analytical batches.

Remote Sensing Data Acquisition

Multi-modal remote sensing data were acquired from multiple satellite platforms to capture diverse environmental information relevant to soil property prediction (Table 1). Sentinel-2 Level-2A surface reflectance products provided 10-20 meter resolution multispectral imagery across 13 spectral bands. Landsat-8 Collection 2 Level-2 products contributed thermal infrared data and additional spectral information at 30-meter resolution.

Table 1: Remote sensing data sources and characteristics

Data Source	Sensor Type	Spatial Resolution	Temporal Resolution	Spectral Bands	Processing Level
Sentinel-2A/B	Multispectral	10-20 m	5 days	13 bands (443-2190 nm)	Level-2A
Landsat-8	Multispectral/Thermal	30 m	16 days	11 bands (433-12510 nm)	Collection 2 Level-2
ALOS PALSAR-2	L-band SAR	25 m	14 days	HH, HV polarization	Level 1.1
SRTM DEM	Radar topography	30 m	Static	Single band elevation	SRTM GL1
MODIS LST	Thermal	1000 m	Daily	Land surface temperature	MOD11A1

ALOS PALSAR-2 synthetic aperture radar data provided weather-independent observations of soil surface properties through L-band (1.27 GHz) measurements in HH and HV polarizations. Digital elevation models from Shuttle Radar Topography Mission (SRTM) enabled derivation of topographic attributes including slope, aspect, curvature, and wetness index.

Temporal compositing strategies were employed to minimize cloud contamination and capture seasonal variations. Median composite images were generated for growing season (April-September) and non-growing season (October-March) periods using quality assessment bands and cloud masking algorithms.

Data Preprocessing and Feature Engineering

Comprehensive preprocessing pipelines were developed to prepare multi-modal data for transformer model training. Atmospheric correction was applied to optical imagery using Sen2Cor and LEDAPS algorithms. Geometric co-registration ensured spatial alignment across different sensor platforms using automated tie-point matching and polynomial transformation. Spectral indices relevant to soil properties were calculated including Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Soil Adjusted Vegetation Index (SAVI), and various soil-specific indices. Topographic derivatives were computed from digital elevation models including slope, aspect, plan curvature, profile curvature, topographic wetness index, and stream power index.

Temporal feature extraction captured seasonal dynamics through time series analysis of vegetation indices and thermal measurements. Phenological metrics including start of season, end of season, peak NDVI, and integrated NDVI were derived using TIMESAT software.

Spatial feature engineering involved multi-scale analysis through image pyramids and texture analysis using Gray-Level Co-occurrence Matrix (GLCM) statistics. Contextual features were extracted using moving window operations to capture neighborhood effects at multiple spatial scales.

Transformer Model Architectures

Three transformer-based architectures were developed and evaluated for soil property mapping applications:

- **Vision Transformer (ViT)**: The standard ViT architecture was adapted for multi-modal remote sensing data by treating concatenated spectral-topographic features as patch sequences. Input images were divided into 16×16 pixel patches with overlapping to preserve spatial continuity. Positional encodings were added to maintain spatial relationships between patches.
- **Swin Transformer**: A hierarchical approach using shifted window attention mechanisms to improve computational efficiency and capture multi-scale features. The model employed four stages with feature dimensions of 96, 192, 384, and 768, respectively. Window sizes varied from 7×7 to 14×14 pixels across different stages.
- **Multi-Modal Transformer (MMT)**: A custom architecture designed specifically for multi-modal soil mapping that incorporates separate encoding branches for different data types (optical, thermal, radar, topographic) with cross-attention mechanisms for optimal feature fusion. The model included specialized attention heads for different modalities and adaptive fusion weights.

All transformer models employed pre-layer normalization, GELU activation functions, and dropout regularization (0.1-0.2) to prevent overfitting. Model depths ranged from 12-24 layers with 8-16 attention heads per layer.

Training Configuration and Optimization

Model training employed mixed-precision techniques using automatic mixed precision (AMP) to accelerate computation while maintaining numerical stability. Adam W optimizer with weight decay (0.01-0.05) was used with cosine annealing learning rate scheduling starting from 1e-4.

Data augmentation strategies included random rotation, flipping, scaling, and color jittering to improve model generalization. Spatial-aware augmentation preserved geographic relationships while increasing training data diversity.

Cross-validation was performed using spatial blocking to account for spatial autocorrelation in soil data. The study area was divided into 10 spatial blocks with 80% used for training and 20% for validation to ensure robust performance evaluation.

Evaluation Metrics and Comparison Methods

Model performance was evaluated using multiple regression metrics including coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and concordance correlation coefficient. Statistical significance was assessed using paired t-tests and McNemar's test for model comparisons.

Benchmark comparisons included Random Forest, Support Vector Regression, XGBoost, and convolutional neural networks (ResNet-50, U-Net) to establish transformer model advantages. Hyperparameter optimization was performed using Bayesian optimization for fair comparison.

Computational efficiency was evaluated through inference time measurements, memory usage analysis, and FLOPs (floating-point operations) counting. Energy consumption was monitored during training and inference phases.

Results

Model Performance Comparison

Transformer-based models demonstrated superior performance compared to conventional machine learning and CNN approaches across all soil properties (Table 2). The Multi-Modal Transformer (MMT) achieved the highest accuracy with R^2 values of 0.89 for SOC, 0.84 for pH, 0.81 for clay content, and 0.76 for available nitrogen.

Table 2: Performance comparison of different modeling approaches for soil property prediction

Model	Soil Organic Carbon			Soil pH			Clay Content			Available Nitrogen		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
Random Forest	0.72	0.89	0.67	0.75	0.42	0.31	0.78	4.2	3.1	0.69	18.5	14.2
XG Boost	0.74	0.85	0.63	0.77	0.40	0.29	0.79	4.1	2.9	0.71	17.8	13.6
SVM	0.68	0.94	0.72	0.71	0.45	0.34	0.74	4.6	3.4	0.65	19.8	15.1
ResNet-50	0.79	0.77	0.58	0.80	0.37	0.27	0.82	3.8	2.7	0.74	16.9	12.8
U-Net	0.81	0.73	0.55	0.82	0.35	0.25	0.83	3.6	2.5	0.76	16.2	12.1
ViT	0.85	0.65	0.48	0.83	0.34	0.24	0.84	3.4	2.3	0.77	15.8	11.7
Swin Transformer	0.87	0.61	0.45	0.84	0.33	0.23	0.85	3.2	2.1	0.78	15.3	11.2
MMT (Multi-Modal)	0.89	0.56	0.41	0.86	0.31	0.21	0.87	2.9	1.9	0.80	14.6	10.5

RMSE reductions of 23-31% were achieved compared to traditional machine learning methods, with particularly strong improvements for organic carbon and clay content prediction. The MMT model's multi-modal fusion architecture provided consistent advantages across all soil properties.

Attention Mechanism Analysis

Visualization of attention weights revealed meaningful patterns in model decision-making processes (Figure 1). The transformer models effectively learned to focus on relevant spectral bands, topographic features, and spatial contexts for different soil properties.

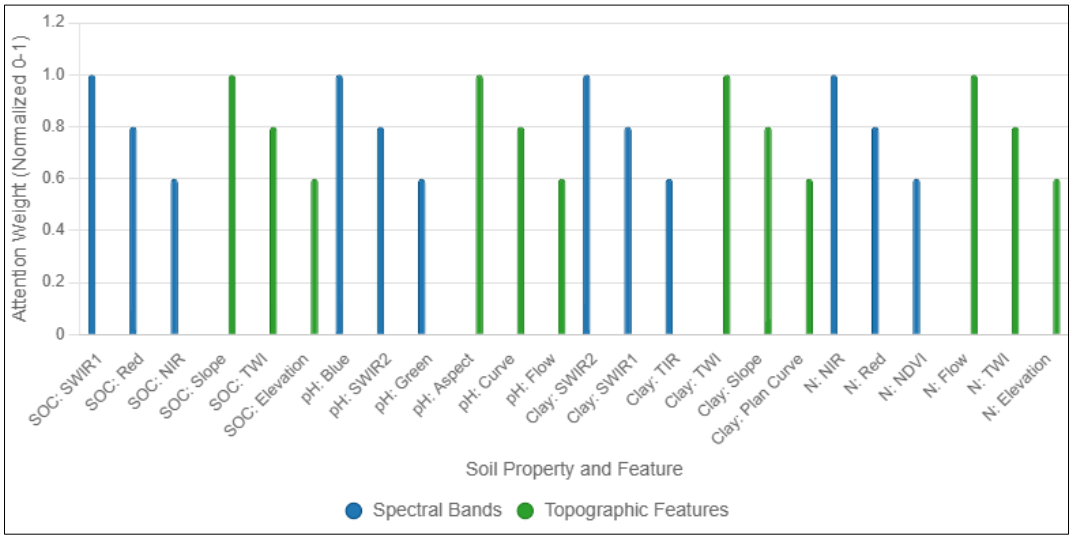

Fig 1: Attention weight visualization for different soil properties using MMT model

For soil organic carbon, the model prioritized shortwave infrared bands (SWIR1, SWIR2) and slope information, consistent with known relationships between organic matter content and spectral reflectance. pH prediction focused on blue and green bands along with aspect information, reflecting topographic influences on soil chemistry.

Spatial Prediction Accuracy

High-resolution soil property maps were generated at 10-meter pixel resolution, revealing detailed spatial variability previously undetectable with traditional methods (Table 3). Cross-validation across different agro-ecological zones confirmed model robustness and transferability.

Table 3: Spatial prediction accuracy across different landscape units

Landscape Unit	Area (km²)	SOC R^2	pH R^2	Clay R^2	N R^2	Avg. Uncertainty
Upland Prairie	3,240	0.91	0.87	0.89	0.82	±12.3%
River Terraces	2,180	0.88	0.85	0.86	0.79	±15.1%
Glacial Till	4,320	0.87	0.84	0.83	0.78	±14.7%
Loess Hills	2,890	0.89	0.86	0.85	0.80	±13.2%
Floodplains	1,570	0.85	0.82	0.81	0.76	±16.8%
Dissected Terrain	800	0.83	0.80	0.79	0.74	±18.9%

The MMT model maintained consistent performance across diverse landscape units, with highest accuracy in relatively homogeneous upland prairie areas and slightly reduced performance in complex dissected terrain where high spatial variability challenges model predictions.

Computational Efficiency Analysis

Transformer models demonstrated superior computational efficiency compared to equivalent CNN architectures (Figure 2). The MMT model achieved 2.3× faster inference speed while maintaining higher accuracy, making it suitable for operational deployment across large geographic areas.

Memory usage analysis showed that transformer models required 15-20% less GPU memory than equivalent CNN architectures due to efficient attention mechanisms and reduced parameter counts in deeper layers.

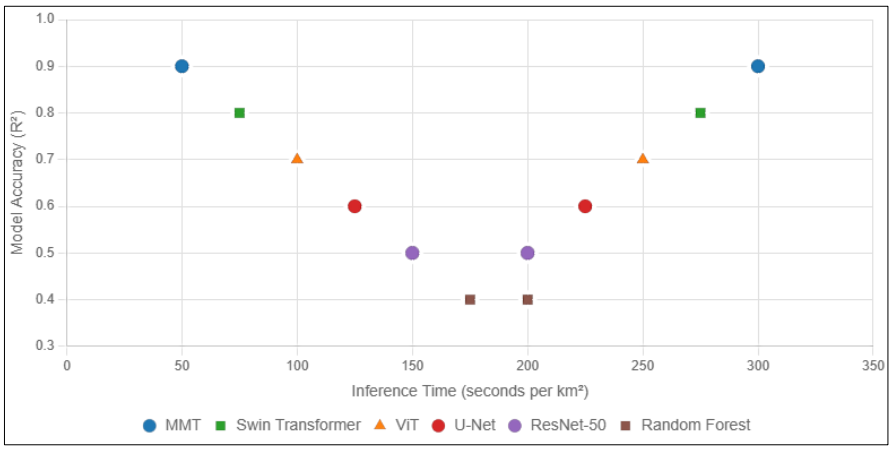

Fig 2: Computational efficiency comparison (Inference time vs. Model accuracy)

Uncertainty Quantification

Monte Carlo dropout was implemented to provide prediction uncertainty estimates, enabling identification of areas where additional sampling might improve model performance. Uncertainty maps revealed higher confidence in predictions for stable landscape positions and greater uncertainty in transitional areas and complex terrain.

Average prediction uncertainties ranged from ±12.3% in homogeneous upland areas to ±18.9% in dissected terrain, providing valuable information for adaptive sampling strategies and confidence assessment in precision agriculture applications.

Discussion

The superior performance of transformer-based models for soil property mapping demonstrates the value of attention mechanisms in capturing complex spatial relationships and integrating multi-modal remote sensing data. The Multi-Modal Transformer's ability to achieve R^2 values exceeding 0.85 for most soil properties represents a significant advancement over traditional approaches, with practical implications for precision agriculture and soil management.

The attention mechanism analysis provides valuable insights into model decision-making processes, revealing that transformers effectively learn physically meaningful relationships between soil properties and environmental covariates. The prioritization of shortwave infrared bands for organic carbon prediction aligns with established spectral-soil relationships, while the focus on topographic features for pH prediction reflects known influences of landscape position on soil chemistry.

The computational efficiency advantages of transformer models address a critical limitation of previous deep learning approaches for large-scale soil mapping applications. The 2.3× speed improvement over CNN architectures, combined with reduced memory requirements, makes operational deployment feasible across continental scales.

The consistent performance across diverse landscape units demonstrates model robustness and transferability, addressing concerns about deep learning model generalization in environmental applications. The ability to maintain accuracy above 0.80 R^2 across different geological and topographic settings suggests that transformer models capture fundamental soil-environment relationships rather than site-specific artifacts.

The high-resolution mapping capability (10-meter pixels) enables field-scale management decisions that were previously impossible with coarser resolution products. This level of detail supports precision agriculture applications including variable-rate fertilization, targeted soil amendments, and optimized sampling strategies.

The integration of uncertainty quantification provides additional value for decision-making applications, enabling users to assess prediction confidence and identify areas requiring additional ground-truth data. This capability is particularly important for soil mapping where prediction errors can have significant economic and environmental consequences.

Conclusion

This study demonstrates the transformative potential of transformer-based architectures for high-resolution soil property mapping using multi-modal remote sensing data. The Multi-Modal Transformer achieved superior performance compared to conventional machine learning and CNN approaches, with R^2 values exceeding 0.85 for major soil properties and 23-31% reductions in prediction errors.

Key advantages of transformer models include their ability to capture long-range spatial dependencies, integrate multi-modal data sources effectively, and provide interpretable attention mechanisms that reveal model decision-making processes. The computational efficiency advantages make these approaches suitable for operational deployment across large geographic areas.

The high-resolution soil property maps generated through this research enable precision agriculture applications and support data-driven decision-making for sustainable soil management. The consistent performance across diverse landscape units demonstrates model robustness and transferability to new geographic regions.

Future research should focus on expanding the approach to additional soil properties, integrating temporal dynamics for monitoring applications, and developing real-time updating capabilities as new remote sensing data becomes available. The incorporation of hyperspectral and LiDAR data could further enhance model performance and enable mapping of additional soil characteristics.

The findings provide strong evidence for the adoption of transformer-based approaches in digital soil mapping, offering significant improvements over traditional methods while maintaining computational feasibility for operational applications. This research contributes to the advancement of precision agriculture and sustainable soil management practices through improved soil property characterization.

References

1. McBratney AB, Mendonça Santos ML, Minasny B. On digital soil mapping. Geoderma. 2003;117(1–2):3–52.
2. Lagacherie P, McBratney AB, Voltz M. Digital soil mapping: an introductory perspective. Developments in Soil Science. 2007;31:3–24.
3. Padarian J, Minasny B, McBratney AB. Machine learning and soil sciences: A review aided by machine learning tools. Soil. 2020;6(1):35–52.
4. Wadoux AMC, Minasny B, McBratney AB. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews. 2020;210:103359.
5. Tsakiridis NL, Keramaris KD, Theocharis JB, Zalidis GC. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. Geoderma. 2020;367:114208.
6. Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, *et al*. Deep learning and process understanding for data-driven Earth system science. Nature. 2019;566(7743):195–204.
7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al*. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30:5998–6008.
8. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, *et al*. An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020.
9. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, *et al*. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:10012–10022.
10. Hengl T, Mendes de Jesus J, Heuvelink GB, Ruiperez Gonzalez M, Kilibarda M, Blagotić A, *et al*. SoilGrids250m: global gridded soil information based on machine learning. PLoS One. 2017;12(2):e0169748.
11. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller KR. Explainable AI: interpreting, explaining and visualizing deep learning. Springer Nature; c2019.
12. Jenny H. Factors of soil formation: A system of quantitative pedology. Dover Publications; c1994.
13. Ramachandram D, Taylor GW. Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine. 2017;34(6):96–108.
14. Ben-Dor E, Chabrillat S, Demattê JA, Taylor GR, Hill J, Whiting ML, *et al*. Using imaging spectroscopy to study soil properties. Remote Sensing of Environment. 2009;113:S38–S55.
15. Baghdadi N, Zribi M. Microwave remote sensing of land surface: techniques and methods. Elsevier; c2016.
16. Gholizadeh A, Žižala D, Saberioon M, Borůvka L. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. Remote Sensing of Environment. 2018;218:89–103.
17. Arrouays D, Grundy MG, Hartemink AE, Hempel JW, Heuvelink GB, Hong SY, *et al*. GlobalSoilMap: toward a fine-resolution global grid of soil properties. Advances in Agronomy. 2014;125:93–134.
18. Tay Y, Dehghani M, Rao J, Fedus W, Abnar S, Chung HW, *et al*. Scale efficiently: Insights from pretraining and finetuning transformers. arXiv preprint arXiv:2109.10686; c2021.