remote sensing

MDPI

Article

Transformer-Guided Noise Detection and Correction in Remote Sensing Data for Enhanced Soil Organic Carbon Estimation

Manoranjan Paul [1,2], Dristi Datta[1,2,*], Manzur Murshed [3] , Shyh Wei Teng [4] and Leigh M. Schmidtke [5]

[1] School of Computing, Mathematics, and Engineering, Charles Sturt University, Bathurst, NSW 2795, Australia; mpaul@csu.edu.au
[2] Cooperative Research Centre for High Performance Soils, Callaghan, NSW 2308, Australia
[3] School of Information Technology, Deakin University, Burwood, VIC 3125, Australia; m.murshed@deakin.edu.au
[4] Institute of Innovation, Science and Sustainability, Federation University Australia, Berwick, VIC 3806, Australia; shyh.wei.teng@federation.edu.au
[5] Gulbali Institue, Charles Sturt University, Wagga Wagga, NSW 2650, Australia; lschmidtke@csu.edu.au
* Correspondence: ddatta@csu.edu.au

Highlights

What are the main findings?

- A unified noise detection–correction framework that learns spectral representations with a Transformer, detects noisy samples via Isolation Forest, and corrects reflectance with a cGAN for SOC estimation.
- Outperforms existing noise-handling methods on benchmark satellite datasets.

What is the implication of the main finding?

- Correcting (rather than discarding) noisy samples improves accuracy and coverage of SOC maps at scale.
- Enables reliable remote sensing–based SOC monitoring to support precision agriculture and soil management.

check for updates

Academic Editor: Fubo Zhao

Received: 7 September 2025
Revised: 14 October 2025
Accepted: 15 October 2025
Published: 17 October 2025

Citation: Paul, M.; Datta, D.; Murshed, M.; Teng, S.W.; Schmidtke, L.M. Transformer-Guided Noise Detection and Correction in Remote Sensing Data for Enhanced Soil Organic Carbon Estimation. *Remote Sens.* **2025**, *17*, 3463. https://doi.org/10.3390/rs17203463

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

Abstract

Soil organic carbon (SOC) is a critical indicator of soil health, directly influencing crop productivity, soil structure, and environmental sustainability. Existing SOC estimation techniques using satellite reflectance data are effective for large-scale applications; however, their accuracy is reduced due to various types of noisy samples caused by vegetation interference, sensor-related anomalies, atmospheric effects, and other spectral distortions. This study proposes a robust data refinement framework capable of handling any soil sample, whether clean or noisy, by identifying and correcting noisy samples to enable more accurate SOC estimation outcomes. The approach first explores complex global relationships among spectral bands to understand and represent subtle patterns in soil reflectance using the Transformer network. To remove redundancy and retain only essential information of the transformed features, we apply a dimensional reduction technique for efficient analysis. Building upon this refined representation, noisy samples are detected without relying on strict data distribution assumptions, ensuring effective identification of noisy samples in diverse conditions. Finally, instead of excluding these noisy samples, the proposed framework corrects their reflectance through a conditional Generative Adversarial Network (cGAN) to align with expected soil spectral characteristics, thereby preserving valuable information for more accurate SOC estimation. The proposed approach was evaluated on benchmark satellite datasets, demonstrating superior performance over existing noise

correction techniques. Experimental validation using the Landsat 8 dataset demonstrated that the proposed framework improved SOC estimation performance by increasing R^2 by 1.52%, reducing RMSE by 4.45%, and increasing RPD by 5.14% compared to the best baseline method (OC-SVM + Kriging). These results confirm the framework's effectiveness in enhancing SOC estimation under noisy conditions. This scalable framework supports accurate SOC monitoring across diverse conditions, enabling informed soil management and advancing precision agriculture.

Keywords: transformer model; remote sensing; LUCAS data; noise detection; noise correction; conditional GAN

1. Introduction

Accurate estimation of soil organic carbon (SOC) is essential for maintaining soil fertility, supporting sustainable agriculture, and regulating the global carbon cycle as a major carbon sink [1,2]. Traditional laboratory-based SOC measurements, although precise, are labor-intensive, costly, and lack spatial scalability [3,4]. While hyperspectral imaging (HSI) combines high spatial and spectral resolution and has shown promising results for SOC estimation [5–12], its high cost and limited accessibility hinder widespread adoption, especially for large-scale assessments. Visual band multispectral imaging offers a lower-cost alternative with comparable accuracy in localized settings [13–16]; however, it is constrained by its focus on specific areas. To overcome these limitations, satellite remote sensing has been increasingly used for large-scale SOC estimation, enabling cost-effective analysis across vast landscapes.

Remote sensing multispectral data from satellites such as Sentinel-2 (S2) and Landsat 8 (L8) have become prominent tools for large-scale SOC estimation due to their extensive spatial coverage, public availability, and cost-effectiveness for measuring soil properties across landscapes. Recent global initiatives have further highlighted the utility of harmonized satellite datasets for Earth observation in large-scale soil property assessments [17]. However, these satellite datasets are highly sensitive to noise and are influenced by factors such as atmospheric conditions, vegetation cover, and sensor limitations [18–20]. Therefore, extensive preprocessing is required to mitigate noise and maintain data integrity and usability for accurate SOC estimation.

Various approaches have been explored to refine satellite data and mitigate noise effects for SOC estimation. Threshold-based exclusion methods, such as those using the normalized difference vegetation index (NDVI), are commonly employed to identify and exclude vegetation-contaminated pixels, thereby ensuring more accurate spectral data for analysis [19,21–23]. While fixed NDVI thresholds are widely used, recent studies have also explored dynamic or adaptive thresholding approaches that adjust based on local vegetation conditions to improve exclusion accuracy [24]. Additionally, other vegetation indices (VIs), such as the enhanced vegetation index (EVI) and soil-adjusted vegetation index (SAVI), have been used to reduce vegetation impacts on reflectance data. Soil indices (SIs) and Tasseled Cap Transformation (TCT) components have also been incorporated alongside spectral bands to enhance the performance of learning-based SOC estimation models [19,25]. Although these derived features improve the representation of soil and vegetation dynamics, exclusion-based methods, whether fixed or dynamic, primarily remove noisy samples rather than correcting them, which may limit the availability of usable data for large-scale SOC estimation.

Preprocessing of satellite data for SOC estimation often involves detecting and correcting noise to ensure reliable model performance. Various noise detection and correction

methods have been explored in remote sensing studies. Statistical approaches, such as Z-Score and Median Absolute Deviation (MAD), offer robust anomaly detection frameworks, and when combined with geostatistical techniques like Kriging, they can moderately enhance data quality for SOC estimation [26–29]. However, these methods typically assume Gaussian data distributions, limiting their effectiveness for non-Gaussian noise patterns common in satellite data.

Machine learning (ML)-based methods, such as Local Outlier Factor (LOF) and One-Class Support Vector Machine (OC-SVM), provide more flexible anomaly detection by leveraging local or global data structures [30–33]. While these approaches improve noise detection, subsequent correction often relies on reconstruction techniques like Robust PCA or Kriging, which may not fully restore data integrity in high-dimensional and complex remote sensing datasets [34–36].

Deep learning (DL)-based approaches, such as Variational Autoencoders (VAEs) and Conditional VAEs (cVAEs), have recently been explored for noise detection and correction in remote sensing applications due to their ability to model non-linear patterns [37–39]. However, their performance is often dependent on the availability of large and high-quality training datasets, which can be a limitation in certain SOC estimation scenarios.

Despite these advancements, existing noise correction methods predominantly focus on excluding noisy data or partially correcting anomalies, potentially leading to information loss and reduced estimation accuracy. Therefore, there is a need for advanced methodologies capable of effectively detecting and correcting noisy satellite reflectance data to fully leverage available datasets for accurate SOC estimation.

To address these limitations, this study proposes a novel methodology that integrates Transformer-based feature extraction to exploit global correlations among different spectral bands, Principal Component Analysis (PCA) for dimensionality reduction, Isolation Forest for identifying noisy samples, and Conditional Generative Adversarial Networks (cGAN) for correcting noisy samples. Each component of this pipeline addresses specific limitations of existing methods, enabling effective noise handling for SOC estimation with limited soil samples.

SOC estimation from satellite reflectance requires capturing complex, non-local dependencies across spectral bands to distinguish subtle noise from true signal variations. Traditional models often struggle to extract such global relationships effectively. To address this, we employ the Transformer network, whose self-attention mechanism dynamically learns inter-band correlations and produces high-level feature representations. This enables robust differentiation between noise and meaningful spectral patterns. Originally developed for natural language processing tasks [40], Transformers have recently been adapted to remote sensing applications, where they demonstrate superior feature extraction capabilities compared to conventional methods [41].

However, Transformer-derived features are often high-dimensional, introducing redundancy and increasing computational costs [42]. To address this, PCA is applied to reduce dimensionality while retaining the most informative components, resulting in a compact yet discriminative feature set that enhances subsequent model efficiency and generalizability.

For noisy sample detection, Isolation Forest is used due to its ability to efficiently identify anomalies in high-dimensional and imbalanced datasets typical of remote sensing [43]. Unlike density-based or statistical methods that rely on distribution assumptions, Isolation Forest isolates anomalies using tree structures, making it effective for detecting noisy samples in complex SOC datasets [44,45].

Once noisy samples are identified, the proposed cGAN reconstructs their reflectance values to align with the distribution of clean data. Unlike generic applications of cGANs,

our approach leverages a tailored training strategy by pairing peer samples with noisy reflectance profiles and corresponding clean (bare soil) samples from the same region or with similar soil characteristics. This conditional setup enables the model to learn the spectral transformation from noise-contaminated to clean soil reflectance more effectively. By focusing the training on realistic and context-aware mappings, the cGAN generates high-fidelity spectral corrections that preserve essential soil information and enhance SOC estimation accuracy. Conditional GANs have shown strong potential in remote sensing for generating realistic and noise-corrected data [46,47], and our implementation extends this capability through a novel data pairing scheme tailored for soil reflectance correction.

By integrating these components, the proposed methodology effectively detects and corrects noisy samples in satellite data, preserving valuable information and enhancing SOC estimation accuracy. This integrated approach addresses the limitations of traditional exclusion-based methods and existing correction techniques, providing a robust and scalable solution for large-scale SOC estimation.

Experiments conducted on SOC datasets derived from L8 imagery validate the effectiveness of the proposed approach. The methodology consistently outperforms traditional models trained on raw or noise-free datasets, demonstrating superior SOC estimation performance. Furthermore, an analysis of varying noise ratios underscores the robustness of the proposed framework, maintaining high accuracy even with significant noise levels.

To further assess the generalizability of the proposed approach, additional experiments were conducted using an S2 dataset characterized by high vegetation cover. These experiments evaluated how effectively the proposed model handles vegetation noise, a critical challenge in SOC estimation from remote sensing data. The results demonstrate that the framework can mitigate the effects of vegetation noise, ensuring reliable SOC estimations even in complex datasets and across multiple sensor platforms.

To evaluate the effectiveness of the proposed method, we compare it against several relevant and widely adopted noise detection and correction techniques, including both traditional and modern approaches. This comparison aims to highlight the strengths of our framework in accurately identifying and correcting noisy soil reflectance samples while preserving essential information for reliable SOC estimation.

The key contributions of this study are as follows:

- Proposing a novel framework that integrates Transformer-based feature extraction, Isolation Forest for noise detection, and cGAN for noise correction.
- Addressing the limitations of traditional exclusion-based noise-handling methods by correcting noisy samples to enhance data utility.
- Demonstrating the effectiveness and scalability of the proposed methodology across multiple remote sensing platforms, including datasets with complex noise patterns.
- Enhancing SOC estimation accuracy by leveraging advanced feature representation and reconstruction techniques.

This paper is organized as follows: The Dataset Preparation section describes the dataset used in this study. The Proposed Noise Detection Approach section introduces the workflow for identifying and correcting noisy samples using a combination of Transformer networks, Isolation Forest, and cGAN. The Methodology section details the overall framework for SOC estimation, including feature engineering and the integration of ML and DL models. The Results section presents a comprehensive evaluation of model performance under different noise scenarios. The Discussion section delves into the implications of the findings, comparing the proposed method with state-of-the-art techniques and examining its practical applicability in large-scale SOC estimation, as well as directions for future research. Finally, the Conclusion section summarizes the key contributions of this study.

2. Dataset Preparation

2.1. Study Area and Soil Data Collection

The Land Use and Land Cover Survey (LUCAS) was launched by EUROSTAT in 2001 as a foundational program to measure the landscape parameters that are key to the assessment of agricultural and environmental changes [48]. Since the establishment of the LUCAS survey, soil has been systematically sampled at specific intervals, while follow-up investigations have been conducted every three years across all EU member states.

In this study, we used the LUCAS 2018 topsoil database, which provides detailed soil property information based on 18,984 soil samples collected throughout Europe. These properties include pH ($CaCl_2$ and H_2O), electrical conductivity (EC), SOC, carbonate ($CaCO_3$), phosphorus (P), total nitrogen (N), and extractable potassium (K), as recorded in the European Soil Data Center (ESDAC). The `LUCAS-SOIL-2018.shp` file contains individual soil sample locations along with their geospatial coordinates, allowing for precise overlaying with satellite imagery.

For this study, only bare soil samples with minimal vegetation cover were selected, ensuring a more accurate assessment of soil reflectance properties. The selection criterion was based on NDVI values, where samples with $0 < \text{NDVI} < 0.30$ were considered to have minimal vegetation interference. The spatial distribution of these selected soil sample points is shown in Figure 1. The dataset includes soil samples taken at depths of 0 to 20 cm, representing the topsoil layer crucial for SOC estimation.

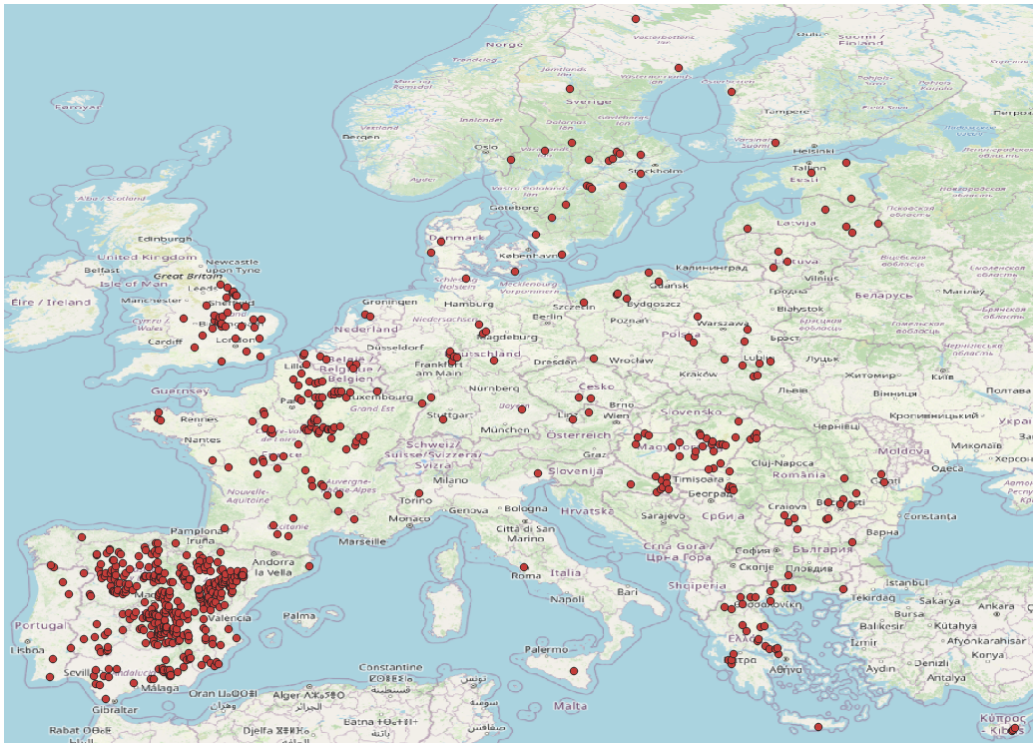


Figure 1. Geographic distribution of LUCAS 2018 bare soil sample locations used in this study. The map shows the spatial extent of sampling across Europe based on the selected dataset.

Table 1 presents the descriptive statistics of the SOC dataset used in this study. SOC concentration is expressed in grams of organic carbon per kilogram of soil (g kg^{-1}), representing the organic carbon content within the 0–20 cm topsoil layer, which is a critical zone for SOC monitoring. In the table, "S. No." indicates the total number of bare soil samples collected, "Min" and "Max" denote the observed range of SOC values, "Mean" and "Median" reflect central tendency, "Std." indicates the standard deviation, and "CV" represents the coefficient of variation (in percentage), highlighting relative variability across samples.

Table 1. Descriptive statistical parameters for the SOC dataset investigated in this study.

Soil Type	NDVI Range	S. No.	Min	Max	Mean	Median	Std.	CV(%)
Bare soil	$0 < \text{NDVI} < 0.30$	510	2.20	96.50	15.05	12.50	11.21	74.48

These statistical attributes emphasize the dataset's variability and heterogeneity under bare soil conditions, which are essential for evaluating the robustness of the proposed noise detection and correction strategy. The large variation in SOC values (ranging from 2.20 to 96.50 g kg^{-1}) reflects the diverse soil properties across different geographic regions, further underscoring the need for a reliable noise-handling framework.

2.2. Landsat 8 Data Acquisition and Pre-Processing

To complement the ground-based soil data with remote sensing information, we acquired corresponding L8 satellite imagery. The L8 data were retrieved from the USGS Earth Explorer platform (https://earthexplorer.usgs.gov/), where satellite images were downloaded for locations matching the soil sample points. The spatial extent of the dataset spans approximately 34.8°N–69.5°N latitude and 8.2°W–33.6°E longitude, covering a broad region across central and western Europe where the LUCAS soil samples were collected. To ensure temporal consistency, the images were selected within a time window of 15 days before or after the soil sampling date, with all acquisitions occurring between May and September 2018 to align with the LUCAS 2018 sampling campaign. A strict cloud cover threshold of less than 10% was applied to minimize atmospheric interference, ensuring high-quality reflectance data for analysis. This time window was chosen to balance the availability of cloud-free images while maintaining reflectance consistency with the soil sampling period.

The band (B) descriptions of L8 are well documented in the literature [49]. For this study, we focused on seven bands (B1–B7) relevant to soil analysis [19], as these bands capture spectral characteristics essential for soil property estimation. We excluded bands primarily intended for atmospheric and thermal analysis (B8–B11) [50]. This selection ensures that only the most relevant spectral information for SOC estimation is retained.

To ensure accurate representation of ground surface conditions, the downloaded L8 images underwent a structured multi-step preprocessing pipeline. First, radiometric calibration was performed to convert raw digital numbers into sensor reflectance values, adjusting for the sensor's position relative to the sun and the time of acquisition [51]. Atmospheric correction was then applied using the Fast Line-of-Sight Atmospheric Analysis of Hypercubes (FLAASH) algorithm, converting top-of-atmosphere reflectance into surface reflectance values to ensure minimal atmospheric interference [52].

To enhance spatial resolution, the pansharpening technique using the Gram–Schmidt algorithm was applied [53]. This method fused the 30-meter resolution multispectral bands (B1–B7) with the higher resolution panchromatic band (B8) at 15 m, improving spatial detail while preserving spectral fidelity. This enhancement is particularly crucial for accurately delineating soil features, ensuring precise extraction of reflectance values.

The preprocessing steps, including radiometric calibration, atmospheric correction, and pansharpening, were performed using ENVI (version 5.6.1). After preprocessing, the L8 images were imported into ArcGIS Pro (version 2.8.0) to extract reflectance values corresponding to the exact locations of the soil sample. By overlaying the processed imagery with geospatial soil sample locations, accurate spectral data were obtained for further analysis. These reflectance values for the seven L8 bands were compiled into an Excel file for subsequent analysis.

2.3. Landsat 8 Image Transformation and Vegetation Indices

To enhance the dataset, several vegetation indices (VIs) and soil indices (SIs) were computed using the reflectance values from the seven L8 bands. These indices are widely used to quantify soil and vegetation properties that influence SOC estimation. The computed VIs include the Ratio Vegetation Index (RVI), NDVI, Green Normalized Difference Vegetation Index (GNDVI), EVI, Soil Adjusted Vegetation Index (SAVI), and Modified Soil Adjusted Vegetation Index (MSAVI). Similarly, the SIs include the Brightness Index (BI), Salinity Index (SI), Color Index (CI), Dry Soil Index (DSI), and Dry Vegetation Index (DVI). VIs primarily help in assessing vegetation cover, while SIs are useful in characterizing soil properties such as texture, moisture, and salinity, all of which have direct implications for SOC distribution. Although the dataset includes only bare soil samples (NDVI < 0.30), vegetation indices were used to capture subtle spectral variations from minimal residues, complementing raw reflectance and enhancing SOC estimation accuracy. The formulas and details of each index are summarized in Table 2, with references to the original works [19,54–56].

Table 2. Vegetation and soil indices used in this study, with corresponding Landsat 8 band-based formulas. Here, *B* represents the reflectance value of Landsat 8 Band.

Index Name	Formula	Source
Ratio Vegetation Index (RVI)	$B5/B4$	[57]
Normalized Difference Vegetation Index (NDVI)	$(B5 - B4)/(B5 + B4)$	[58]
Green NDVI (GNDVI)	$(B5 - B3)/(B5 + B3)$	[54]
Enhanced Vegetation Index (EVI)	$2.5 \times (B5 - B4)/(B5 + 6B4 - 7.5B2 + 1)$	[59]
Soil-Adjusted Vegetation Index (SAVI)	$1.5 \times (B5 - B4)/(B5 + B4 + 0.5)$	[60]
Modified SAVI (MSAVI)	$\frac{2B5+1-\sqrt{(2B5+1)^2-8(B5-B4)}}{2}$	[55]
Brightness Index (BI)	$\sqrt{B4^2 + B3^2}/2$	[56]
Salinity Index (SI)	$\sqrt{B3 \times B4}$	[56]
Color Index (CI)	$(B4 - B3)/(B4 + B3)$	[56]

Similarly, the TCT was performed on the L8 imagery. TCT is a widely used dimensionality reduction technique in remote sensing that projects the spectral bands into new components representing key landscape features, including brightness, greenness, and wetness [61]. This transformation captures crucial soil and vegetation characteristics that may not be explicitly represented by individual spectral bands or traditional VIs. The specific band coefficients used for this transformation are listed in Table 3, based on the formulation in [61]. These derived TCT features were incorporated into the model as additional inputs to support the SOC estimation task.

Table 3. Coefficients for Tasseled Cap Transformation (TCT) components for Landsat 8 bands [61].

TCT Component	B2	B3	B4	B5	B6	B7
Brightness	0.3029	0.2786	0.4733	0.5599	0.5080	0.1872
Greenness	−0.2941	−0.2430	−0.5424	0.7276	0.0713	−0.1608
Wetness	0.1511	0.1973	0.3283	0.3407	−0.7117	−0.4559
TCT4	−0.8239	0.0849	0.4396	−0.0580	0.2013	−0.2773
TCT5	−0.3294	0.0557	0.1056	0.1855	−0.4349	0.8085
TCT6	0.1079	−0.9023	0.4119	0.0575	−0.0259	0.0252

The unified dataset integrates both spectral and spatial attributes, combining VIs, SIs, and TCT with pre-processed L8 reflectance data. By leveraging multiple feature types, this enhanced dataset provides a more comprehensive representation of soil and vegetation dynamics, improving the robustness and accuracy of SOC estimation from remote sensing data.

3. Methodology

3.1. Overview of the Proposed Framework

This study presents an integrated method to detect and correct noisy soil samples from L8 imagery to improve SOC estimation. The proposed framework combines advanced ML and DL techniques, including Transformer-based feature extraction, PCA, Isolation Forest for anomaly detection, and cGAN for noise correction, shown in Figure 2.

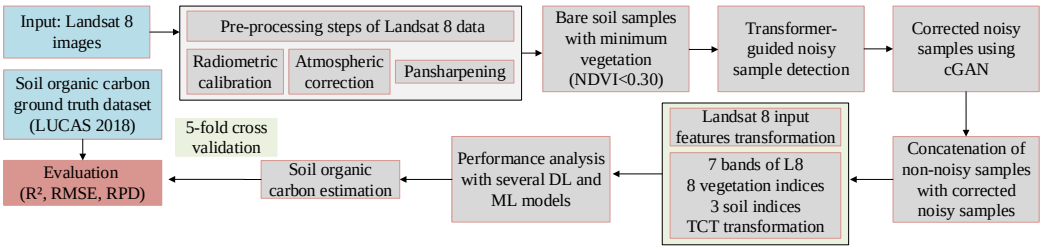


Figure 2. Overview of the methodology for SOC estimation using Landsat 8 imagery, involving noise detection with a Transformer, reconstruction with cGAN, and final SOC estimation using ML and DL models.

The input data, consisting of SOC values and L8 images, were preprocessed to ensure quality and consistency. The preprocessing steps included radiometric calibration, atmospheric correction, and panchromatic sharpening to generate seven spectral bands for analysis. Soil samples with minimal vegetation (NDVI < 0.3) were selected to reduce vegetation noise.

High-level features are extracted using a Transformer encoder and then reduced in dimension using PCA. These hybrid features, along with the original L8 bands, are analyzed using the Isolation Forest algorithm to detect outlier (noisy) samples. Identified noisy samples are corrected using cGAN, which is conditioned on nearby non-noisy sample representations and their average SOC to generate spectrally consistent reflectance values. The corrected samples are then combined with the clean samples to create an augmented dataset, which is used for the downstream task of SOC estimation.

3.2. Proposed Noise Detection and Correction Modules

Identifying and correcting noisy soil samples in satellite reflectance data is critical to improving SOC estimation. This study presents a noise detection and correction approach that combines advanced DL and ML methods. The overall workflow (Figure 3) integrates feature extraction using a Transformer network with PCA for dimensionality reduction, an Isolation Forest for noise identification, and a cGAN for correcting noisy samples by producing clean reflectance. Clean samples are passed directly to the estimation pipeline, while only the identified noisy samples are corrected using the cGAN. By sequentially transforming, filtering, and correcting reflectance data, the workflow ensures that only informative and reliable features are used for SOC estimation. This approach enhances robustness against varying noise levels and maximizes the use of available data. The detailed architecture of the proposed model is presented in Figure 4.

3.2.1. Transformer-Based Feature Extraction

As the first step of the proposed methodology, high-level representations were extracted from the L8 satellite reflectance data using a Transformer network. The Transformer model was selected for its superior ability to capture complex dependencies and interactions between spectral bands, which traditional models often fail to fully exploit.

Unlike conventional approaches, the Transformer model utilizes a self-attention mechanism, which dynamically assigns different weights to spectral bands based on their importance. This enables the model to learn complex relationships among L8 bands and improves its ability to extract meaningful features while suppressing noise. The self-attention mechanism also enhances the model's robustness in handling spectral variations across diverse SOC conditions.

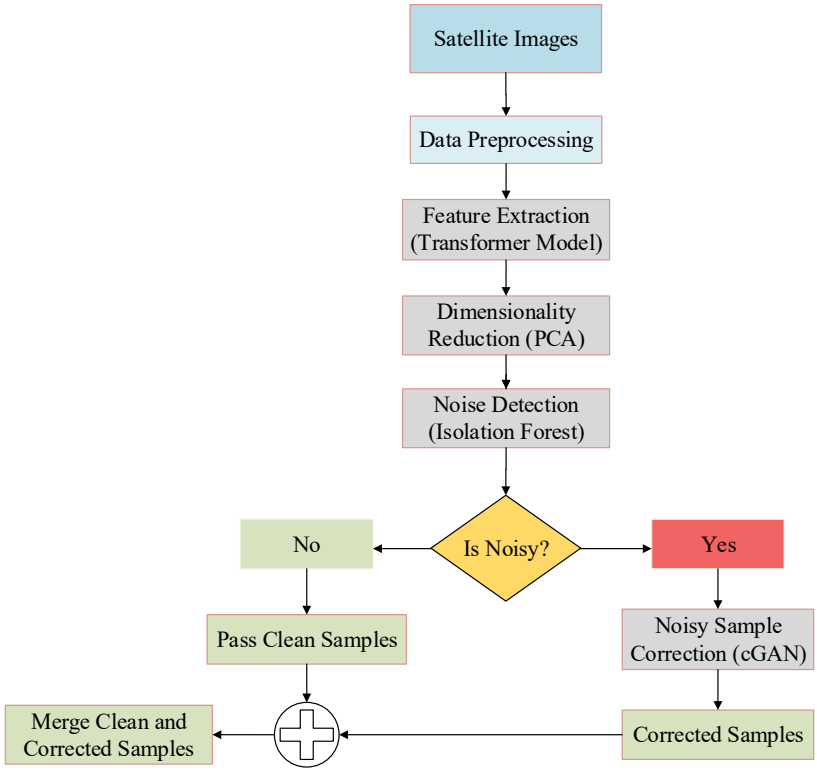


Figure 3. Workflow of the proposed noise detection and correction framework for soil organic carbon estimation. The approach integrates feature extraction, dimensionality reduction, noise detection, and noise correction to identify and correct noisy samples while preserving clean data.

The architectural pipeline of the Transformer-based feature extractor consists of three main components: an embedding layer, a Transformer encoder, and a high-level representation layer. The process begins with a 7-dimensional (7D) input layer, representing the reflectance values of the L8 bands (B1–B7). These raw spectral values are then passed through an embedding layer, where the dimensionality is expanded from 7D to 64D. This embedding step helps the model learn richer feature representations by projecting the input into a higher-dimensional space, making it easier to capture intricate spectral relationships.

The embedded features are then processed by a Transformer encoder, which consists of two layers, each with four self-attention heads. This configuration was chosen empirically to balance model complexity and training stability. The use of multiple self-attention heads allows the model to jointly attend to different inter-band interactions, enabling it to better learn complex spectral dependencies relevant for SOC estimation. The self-attention mechanism allows the model to weigh the contribution of each spectral band dynamically, ensuring that the most relevant information is emphasized while minimizing the influence

of noise. Following the self-attention operation, the output passes through a feedforward network that further refines the extracted features. This structure enables the model to capture both short- and long-range dependencies among spectral bands, improving its ability to distinguish between meaningful spectral variations and noise.

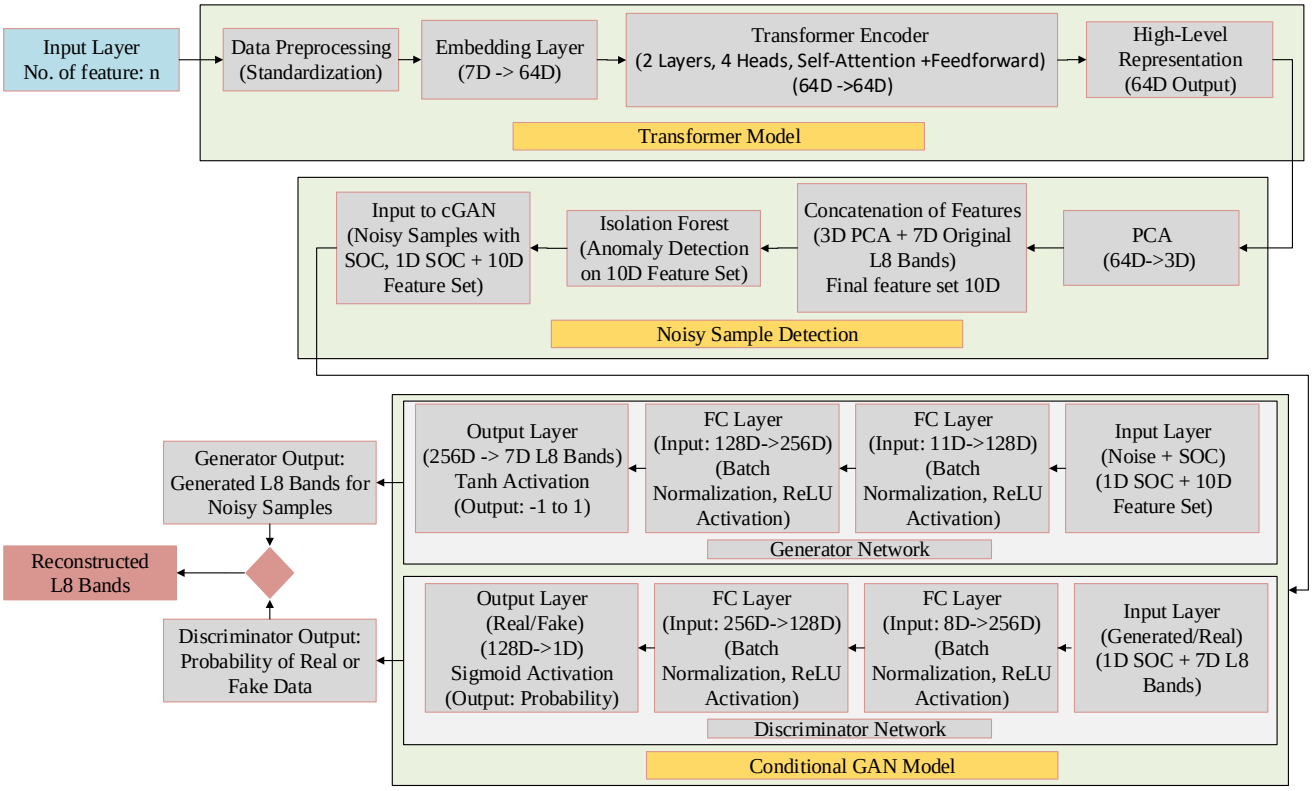


Figure 4. Proposed model architecture for noise detection and correction in SOC estimation, integrating Transformer-based feature extraction, PCA for dimensionality reduction, Isolation Forest for noise detection, and cGAN for correcting noisy samples.

From the L8 seven bands, the Transformer network learns a 64-dimensional feature representation for each soil sample. These high-level features effectively encode the intrinsic relationships between the input bands, facilitating improved noise detection and outlier identification in subsequent stages. By leveraging the self-attention mechanism, the model can automatically highlight influential bands and suppress less informative ones, which is important when dealing with noisy or heterogeneous reflectance data.

3.2.2. Dimensionality Reduction Using Principal Component Analysis

After generating 64-dimensional high-level features using the Transformer, PCA is employed to reduce dimensionality while preserving the most critical information. We selected the first three principal components, which collectively explained 99% of the variance in the dataset. These components capture the most informative spectral variations while reducing noise and redundancy present in the raw data. To construct the final feature set, these three PCA components were combined with the original seven L8 reflectance bands, forming a ten-dimensional feature representation (three PCA components + seven L8 bands). This feature set integrates both the original spectral information and the high-level features extracted by the Transformer, ensuring that the input to the noise detection model is both compact and information-rich.

By leveraging PCA for dimensionality reduction, the methodology enhances the efficiency of noise detection while maintaining the integrity of the spectral data essential for accurate SOC estimation. Furthermore, the transformed feature space provides a more

uniform and informative feature set, which improves the ability of the Isolation Forest model to effectively detect noisy samples within the dataset.

3.2.3. Noise Detection Using Isolation Forest

After creating the feature set, the Isolation Forest algorithm is used to detect noisy samples in the dataset. This unsupervised anomaly detection model isolates outliers by randomly partitioning the data. Since it is based on the assumption that the anomalous points are few and different from most of the data, it effectively identifies noisy samples that deviate from the expected patterns.

The Isolation Forest is trained on the combined feature set (PCA components + original L8 bands) to identify samples that are likely to be noisy. Any noisy samples will be outliers and separated in the next phase for correction. This unsupervised noise detection approach is highly scalable and works well for large satellite datasets by eliminating the need for labeled noisy data.

3.2.4. Noise Correction Using Conditional GAN

After detecting the noisy samples, a cGAN reconstructs their L8 reflectance values. The cGAN model consists of a generator and a discriminator network. The generator produces refined L8 reflectance values for noisy samples, while the discriminator evaluates the authenticity of these generated values by comparing them with non-noisy samples.

Non-noisy samples condition the cGAN, serving as references in the output generation. In this implementation, the generator is conditioned on the measured SOC values of the samples, which guide the reconstruction process by providing soil-property context. This conditioning helps ensure that the generated reflectance remains consistent with the soil's expected characteristics without requiring additional contextual inputs such as spatial neighbors. This approach takes advantage of the observation that SOC typically exhibits limited spatial variability within a specific field [62], ensuring that the generated samples are consistent with the expected range and distribution of SOC in the vicinity. The conditional setup improves the generator's ability to reconstruct plausible reflectance even when noise arises from factors such as sensor artifacts or vegetation contamination.

The generator network takes a concatenated input of noise, SOC values, and the 10-dimensional feature set, passing it through fully connected (FC) layers. The first layer maps the input (11D) to a higher-dimensional space (128D) using batch normalization and ReLU activation. The second layer expands this to 256D, refining the generated spectral reflectance values. Finally, the output layer maps the 256D representation back to 7D, corresponding to the L8 bands, using a Tanh activation function to ensure realistic reflectance scaling.

The discriminator network, in contrast, receives either real or generated L8 reflectance values along with SOC as input (8D). It processes this data through fully connected layers, first mapping it to 256D, followed by 128D, using batch normalization and ReLU activation. The final layer reduces the representation to a single probability score (1D) using a sigmoid activation function, determining whether the input reflectance values are real or synthetic.

For reproducibility, both the generator and discriminator were trained using a composite loss function that balances adversarial realism with direct spectral fidelity. The generator's loss combines an adversarial term (binary cross-entropy) with an L_1 reconstruction penalty to minimize deviations from the true (non-noisy) spectra:

$$\mathcal{L}_G = \mathcal{L}_{adv}^G + \lambda \, \mathcal{L}_{L1},$$

where $\mathcal{L}_{adv}^G$ is the adversarial loss, $\mathcal{L}_{L1} = \|y - \hat{y}\|_1$ is the absolute difference between the real spectrum y and generated spectrum $\hat{y}$, and λ is a weighting factor (empirically

set to 100) to ensure spectral fidelity while still benefiting from adversarial refinement. The discriminator minimizes the standard binary cross-entropy loss $\mathcal{L}_D$ to improve its ability to distinguish real from generated samples.

To improve training stability and prevent overconfidence, label smoothing was applied (real label = 0.9, fake label = 0.1). Optimization was performed using the Adam optimizer with learning rates of 2×10^{-4} for both networks, $\beta_1 = 0.5$, and a batch size of 32. A StepLR scheduler reduced the learning rate by half every 50 epochs. Training was conducted for 100–200 epochs, depending on convergence on the validation set. All hyperparameters were selected based on observed stability and convergence during development.

Several design strategies were incorporated to ensure physical plausibility and statistical consistency of the generated reflectance:

- Conditional guidance: The generator is explicitly conditioned on nearby non-noisy samples and their average SOC, introducing contextual and semantic constraints.
- Direct reconstruction penalty: The L_1 term enforces closeness to true reflectance, reducing large deviations.
- Activation and clipping: Tanh activation followed by inverse normalization ensures outputs fall within a valid physical range (e.g., [0, 1]).
- Standardization: Inputs and outputs were standardized during training and destandardized post-generation to maintain consistency with the original data scale.

This adversarial training loop allows the generator to iteratively minimize the spectral discrepancy between real and generated data, while the discriminator continually improves its classification boundary. This dynamic improves spectral consistency of reconstructed samples and enhances downstream SOC estimation accuracy.

Although the current model was primarily evaluated under typical remote sensing noise scenarios (e.g., vegetation contamination, atmospheric effects), its conditional and data-driven formulation is flexible. The architecture can be easily adapted to other noise types by modifying the conditioning features or augmenting the training set with representative noise examples. For domain-specific use cases, additional constraints—such as a spectral angle-based penalty or structured priors—can be incorporated to further enhance correction performance.

Finally, all hyperparameters and training details are reported for transparency and reproducibility. The full implementation, including training scripts and random seeds, is publicly available in the project repository listed in the Data and Code Availability section.

3.2.5. Post-Reconstruction Dataset

In the last step of our proposed method, we combine the corrected noisy L8 reflectance samples with the non-noisy (clean) samples to form a complete dataset for SOC estimation. This ensures that the SOC estimation models are trained on a clean dataset, diminishing the noise and outliers existing in the previous data.

By reconstructing rather than excluding noisy samples, the proposed framework preserves valuable spectral information, ensuring dataset integrity and minimizing the loss of potentially informative samples. This approach enhances the robustness and accuracy of SOC estimation models by reducing noise effects in satellite data, ultimately leading to more stable estimation. The refined dataset allows learning-based models to generalize better across varying soil conditions, making the methodology more effective for large-scale soil monitoring and carbon estimation.

3.3. Experimental Setup

The final dataset, composed of both clean and corrected samples, is used to train multiple traditional ML and DL models for SOC estimation, including Linear Regression

(LR), Support Vector Regression (SVR), Artificial Neural Networks (ANN), Decision Trees (DT), K-Nearest Neighbors (KNN), Gradient Boosting (GB), Random Forest (RF), CatBoost Regressor (CBR), and a 1D Convolutional Neural Network (1D-CNN).

These models are implemented using the PyTorch (version 1.13.1) and Scikit-learn (version 1.2.2) libraries. Hyperparameter tuning is conducted via five-fold grid search, and the optimized settings are summarized in Table 4.

Table 4. Hyperparameter optimization for all models.

Model	Hyperparameters and Values
LR	Standardization: True; Random state: 42
SVR	Kernel: rbf; C: 1.0; Epsilon: 0.2; Batch size: 32; Epochs: 100; Learning rate: 0.001
ANN	fc1: 128; fc2: 64; fc3: 1; Activation: ReLU; Dropout: 0.3; Batch size: 32; Epochs: 100; Learning rate: 0.001
DT	Random state: 42; Max Depth: None
KNN	Neighbors: 5; Metric: Euclidean
1D-CNN	Conv layers: 3; Filters: 16/32/64; Kernel size: 3; Padding: 1; fc1: 128; Dropout: 0.3; Batch size: 32; Epochs: 100; Learning rate: 0.001
GB	Estimators: 100; Random state: 42
RF	Estimators: 100; Random state: 42
CBR	Learning rate: 0.1; Depth: 10; Loss: RMSE; Iterations: 100

Model performance is evaluated using three standard regression metrics: Coefficient of Determination (R^2), Root Mean Square Error (RMSE), and Ratio of Performance to Deviation (RPD). The mathematical formulations and significance of these metrics have been previously described in the literature [13,63].

4. Results

This section presents the performance of various ML and DL models for estimating SOC using L8 imagery. Table 5 summarizes the regression results for various ML and DL models across different input configurations. Three main configurations were tested: (1) using L8 7 bands without transformation, (2) augmenting L8 bands with transformed features, and (3) applying band selection methods. These configurations include raw, only noise-free, and noise-restored data obtained through the proposed approach.

Table 5. Regression results of ML and DL models in estimating soil organic carbon from Landsat 8 data. Each scenario (SC.1–SC.9) is evaluated with three metrics in three successive rows: (R^2: coefficient of determination, unitless; RMSE: root mean square error in g kg^{-1}; RPD: ratio of performance to deviation, unitless). Higher R^2 and RPD, and lower RMSE, indicate better model performance.

Group	Input	LR	SVR	ANN	DT	KNN	1D CNN	GB	RF	CBR
L8 7 bands	SC.1	0.28	0.39	0.42	0.43	0.53	0.50	0.45	0.55	0.56
		9.11	8.50	8.16	7.98	7.27	7.37	7.57	6.80	6.93
		1.19	1.28	1.33	1.39	1.49	1.49	1.35	1.50	1.50
	SC.2	0.40	0.40	0.50	0.37	0.48	0.55	0.59	0.60	0.62
		7.36	7.59	6.57	7.29	6.79	6.05	5.95	5.63	5.75
		1.32	1.29	1.49	1.41	1.44	1.63	1.66	1.81	1.69
	SC.3 (proposed)	0.40	0.33	0.61	0.32	0.57	0.60	0.57	**0.67**	0.64
		7.67	8.92	6.22	8.08	6.54	6.12	6.65	**5.80**	6.08
		1.37	1.23	1.70	1.32	1.61	1.74	1.62	**1.84**	1.75

Table 5. *Cont.*

Group	Input	LR	SVR	ANN	DT	KNN	1D CNN	GB	RF	CBR
L8 7 bands + transformed features	SC.4	0.32	0.41	0.48	0.53	0.53	0.54	0.56	0.59	0.59
		8.84	8.40	7.34	7.24	6.97	6.78	6.93	6.86	6.82
		1.23	1.30	1.49	1.53	1.56	1.62	1.56	1.58	1.59
	SC.5	0.47	0.41	0.51	0.50	0.53	0.59	0.63	0.56	0.63
		6.75	7.55	6.44	6.50	6.40	5.81	5.64	6.00	5.83
		1.45	1.30	1.53	1.55	1.52	1.70	1.74	1.68	1.67
	SC.6 (proposed)	0.52	0.30	0.59	0.31	0.55	0.56	0.62	0.66	**0.68**
		6.81	9.09	6.28	7.62	6.76	6.40	6.23	5.83	**5.77**
		1.55	1.20	1.68	1.42	1.57	1.66	1.70	1.82	**1.84**
Band selection with state-of-art methods (from SC.6)	SC.7	0.53	0.29	0.61	0.33	0.53	0.50	0.62	0.64	0.65
		6.79	9.14	6.16	7.78	6.88	6.69	6.25	5.96	5.96
		1.56	1.19	1.72	1.37	1.54	1.60	1.69	1.78	1.78
	SC.8	0.40	0.30	0.61	0.23	0.52	0.57	0.61	0.62	0.65
		7.69	9.13	6.17	8.13	6.80	6.34	6.30	6.11	5.90
		1.37	1.20	1.72	1.33	1.56	1.68	1.68	1.73	1.79
	SC.9	0.50	0.31	0.61	0.26	0.56	0.58	0.61	**0.65**	0.65
		6.97	9.01	6.11	7.75	0.57	6.23	6.34	**5.87**	5.95
		1.52	1.21	1.73	1.39	1.61	1.71	1.67	**1.84**	1.80

SC = Scenario; SC.1 = All raw data [13,64], SC.2 = only noise-free data [65,66], SC.3 = Noise-restored data (proposed approach, Figure 5), SC.4 = All raw data with transformed features, SC.5 = Noise-free data with transformed features, SC.6 = Noise-restored data (proposed approach) with transformed features, SC.7 = Lasso selected features from SC.6 (Figure 6a), SC.8 = RFE selected features from SC.6 (Figure 6b), SC.9 = RF selected features from SC.6 (Figure 6c).

4.1. SOC Estimation Using Raw Landsat 8 Reflectance Bands

The results indicate that using unprocessed L8 7-band data, represented by SC.1 (all raw data) and SC.2 (only noise-free data), leads to suboptimal performance across most models. The presence of noise in SC.1 significantly reduces SOC estimation accuracy, whereas SC.2 improves results but at the cost of data exclusion, which may introduce bias. In contrast, SC.3, which applies the proposed noise correction method, demonstrates superior performance. Notably, the RF model achieves the highest accuracy in this group, reflecting a substantial improvement over the baseline results.

Figure 5 presents a comparison of various ML and DL models for SOC estimation in SC.3, highlighting the effectiveness of noise correction. Each subfigure illustrates the relationship between estimated SOC values and actual values, with the red dashed line denoting the ideal 1:1 correspondence. The CBR (Figure 5i) and RF models (Figure 5h) produce results that closely align with this ideal line, particularly in the noise-corrected dataset (SC.3). These findings suggest that noise correction enables models to learn more stable and accurate SOC representations, mitigating distortions introduced by noisy samples. Overall, these results emphasize the importance of noise correction in enhancing model robustness and accuracy in SOC estimation. By restoring valuable spectral information instead of discarding noisy samples, the proposed approach maintains dataset integrity and ensures improved model generalization.

While the results in SC.3 demonstrate the significant performance gains achievable through noise detection and correction alone, they also reveal an inherent limitation: spectral reflectance values, even after correction, may not fully capture subtle variations related to soil properties. To address this, the next stage of analysis integrates transformed features derived from vegetation and soil indices (VIs, SIs) and TCT components, as well as high-level spectral representations from the Transformer–PCA pipeline. These additional

features enhance the descriptive power of the dataset by encoding biophysical and soil-specific characteristics that raw reflectance bands may overlook. This enriched feature space further improves the robustness and precision of SOC estimation, as examined in the following subsection.

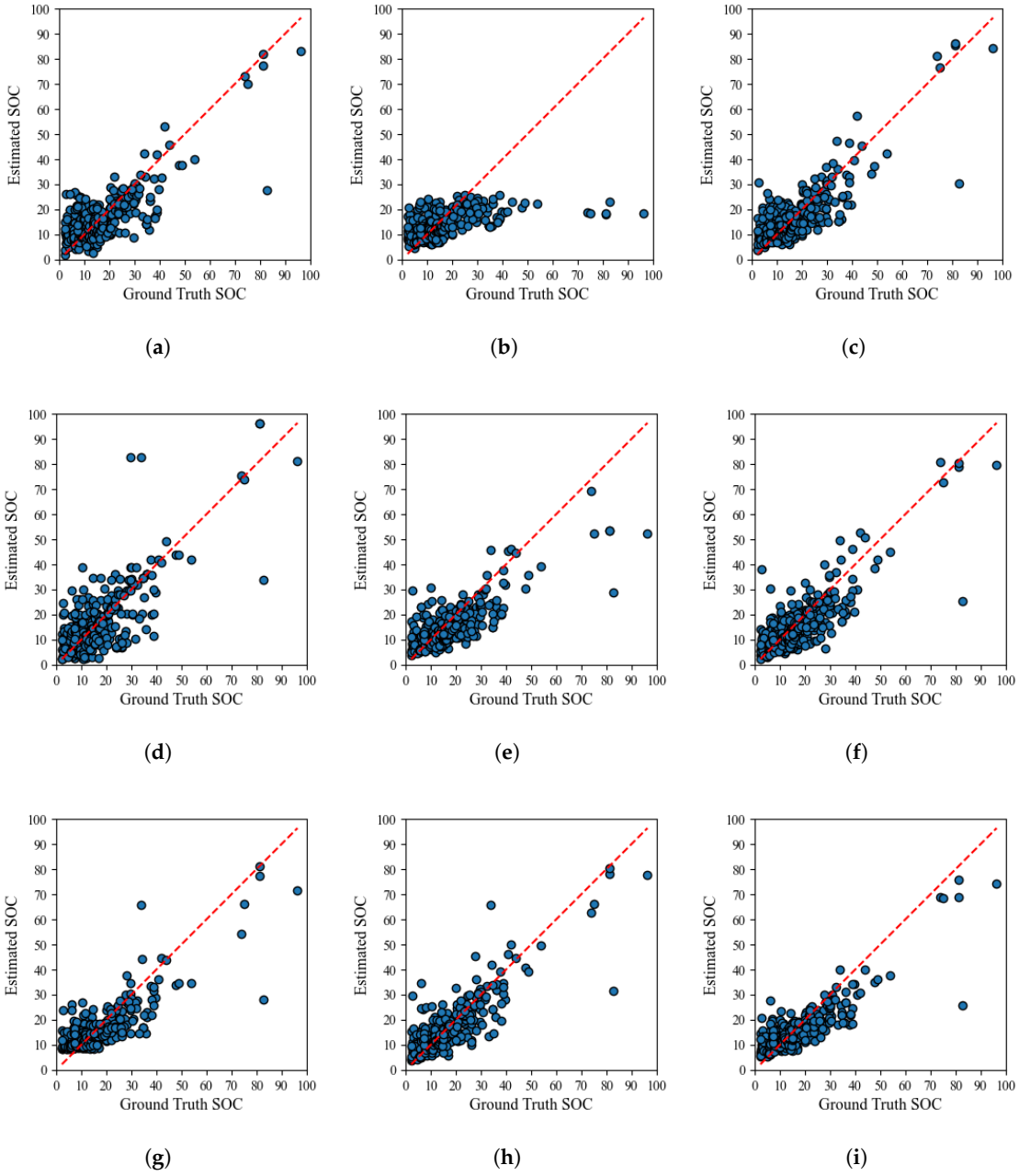


Figure 5. Performance comparison of different ML and DL models for SOC estimation using the SC.3 scenario (proposed approach). Each subfigure shows the model's estimation accuracy, with the red dashed line indicating the 1:1 line where estimated values perfectly match the actual SOC values. Subcaptions include key evaluation metrics: R^2, RMSE, and RPD. (**a**) LR (R^2 = 0.40, RMSE = 7.67, RPD = 1.37). (**b**) SVR (R^2 = 0.33, RMSE = 8.92, RPD = 1.23). (**c**) ANN (R^2 = 0.61, RMSE = 6.22, RPD = 1.70). (**d**) DT (R^2=0.32, RMSE=8.08, RPD=1.32). (**e**) KNN (R^2 = 0.57, RMSE = 6.54, RPD = 1.61). (**f**) 1D-CNN (R^2 = 0.60, RMSE = 6.12, RPD = 1.74). (**g**) GB (R^2 = 0.57, RMSE = 6.65, RPD = 1.62). (**h**) RF (R^2 = **0.67**, RMSE = **5.80**, RPD = **1.84**). (**i**) CBR (R^2 = 0.64, RMSE = 6.08, RPD = 1.75).

4.2. SOC Estimation Using Landsat 8 Bands with Transformed Features: VIs, SIs, and TCT

The second group, which incorporates additional transformed features, shows improved performance across most cases, as presented in Table 5. SC.4 represents transformed features derived from all raw data (SC.1), while SC.5 applies transformations only to noise-free data (SC.2), and SC.6 represents transformed features processed using the proposed approach. These transformations, including VIs, SIs, and TCT components, enhance feature representation and contribute to more stable SOC estimations.

Among these configurations, SC.6, which applies noise correction alongside feature transformations, achieves the best results. This demonstrates that integrating feature transformations with noise reconstruction significantly enhances SOC estimation accuracy. The combined effect of improved spectral features and corrected reflectance values mitigates noise impact while preserving valuable information. Notably, the CBR model shows the most significant improvement in SC.6, as shown in Table 5, reinforcing the advantage of feature transformation alongside noise correction.

4.3. Comparison of Band Selection Techniques for Optimized SOC Estimation

To further enhance model efficiency, we evaluated three widely used band selection techniques: (i) Lasso regression, which applies L1 regularization to shrink coefficients of less informative bands to zero, thereby selecting only the most relevant features; (ii) Recursive Feature Elimination (RFE), which iteratively removes the least significant features based on model weights until the optimal subset is obtained; and (iii) Random Forest (RF) feature importance, which ranks spectral bands by their contribution to reducing estimation error across ensemble decision trees.

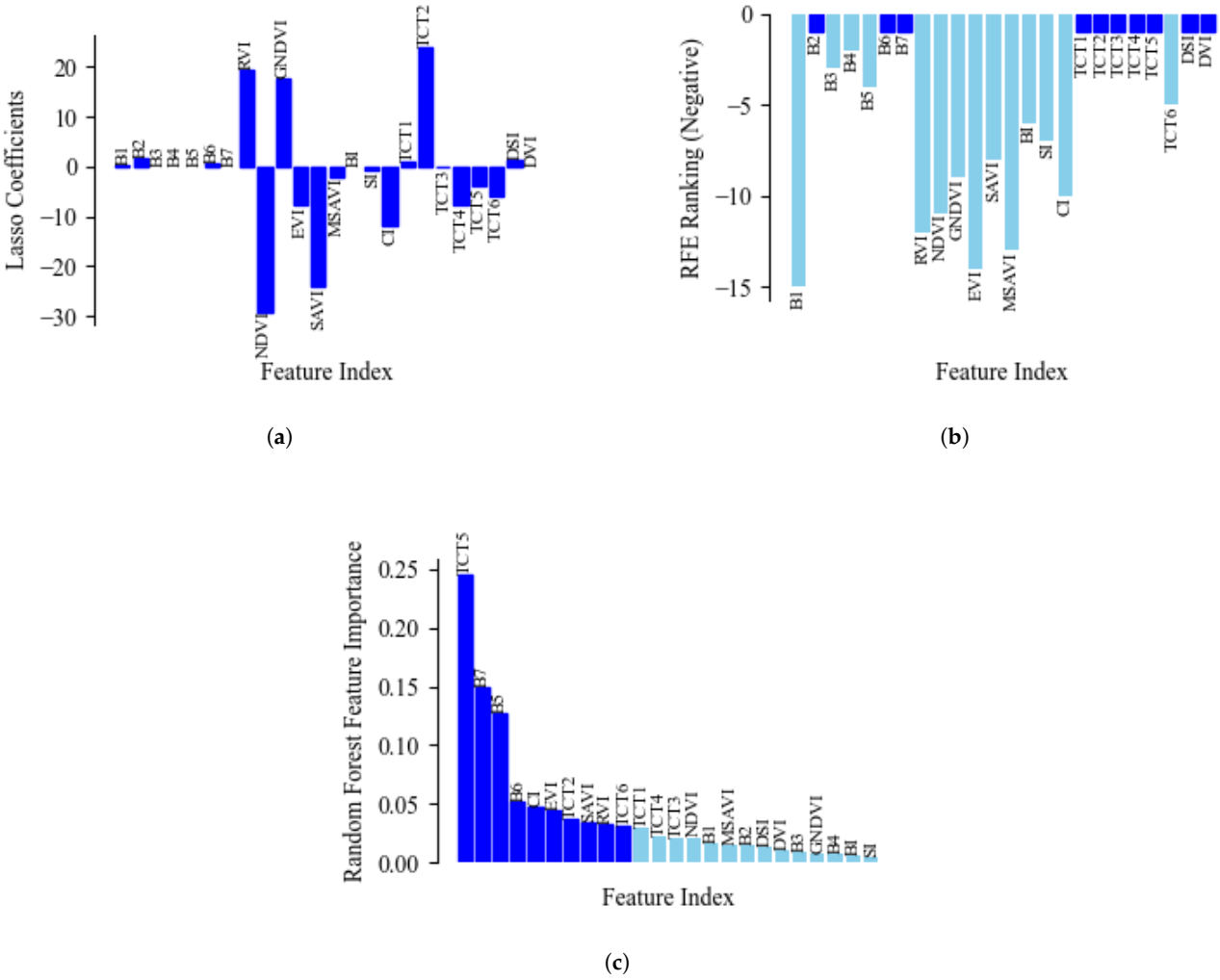
Unlike conventional approaches that apply these techniques directly to raw reflectance data, we apply them to the noise-corrected and feature-enhanced dataset (SC.6). This is a key distinction: by leveraging a cleaned and enriched feature space obtained through the Transformer–PCA pipeline, the band selection process becomes more stable, meaningful, and better aligned with the spectral characteristics relevant to SOC estimation. Scenarios SC.7 to SC.9 (Figure 6) demonstrate how this approach reduces feature complexity and computational time while maintaining high model accuracy.

The results show that while band selection slightly reduces estimation accuracy compared to using all features, it significantly enhances computational efficiency. For instance, in SC.9, RF achieved an R^2 of 0.65, RMSE of 5.87, and RPD of 1.84, highlighting the trade-off between targeted feature selection and model performance. This minor drop in accuracy likely occurs because certain spectral features, though less dominant, still provide valuable information for SOC estimation. However, the substantial gain in computational efficiency makes feature selection highly advantageous for large-scale remote sensing applications.

The conclusion regarding the impact of noise detection and correction is drawn directly from the comparative results across scenarios. As shown in Table 5, noise-corrected datasets (SC.3 and SC.6) consistently yield higher R^2 values and lower RMSE compared with raw (SC.1) and noise-free (SC.2) inputs, demonstrating that removing spectral distortions leads to more accurate SOC estimation. This performance improvement persists even after band selection (e.g., SC.9), indicating that the benefits of noise correction extend throughout the modelling pipeline.

Although the results across different scenarios remain comparable, they underscore the importance of leveraging the full feature set when high precision is required. Overall, these observations confirm that integrating noise detection and correction strengthens model learning across various ML and DL architectures, enhancing SOC estimation performance and reliability. Across all models, CBR and RF consistently demonstrated the highest performance. This suggests that integrating feature selection with noise correction creates

a more optimized and scalable SOC estimation framework, effectively balancing model accuracy and computational efficiency. These results further illustrate the capability of the proposed approach in enhancing SOC estimation from satellite data.



(**a**) (**b**)

(**c**)

Figure 6. Comparison of feature selection results across different algorithms applied to the reconstructed dataset (SC.6), highlighting the importance of selected features in blue. (**a**) Lasso coefficients vs. feature index. (**b**) RFE ranking vs. feature index. (**c**) RF feature importance vs. feature index.

5. Discussion

This section assesses the performance of the proposed noise handling methodology in the estimation of SOC against the state-of-the-art techniques and also analyzes its performance under different noise ratios. Moreover, the versatility of the method is evaluated with an additional S2 dataset, which shows its capability for wider applications. Practical implications, limitations, and directions for future research are also discussed.

5.1. Comparison with State-of-the-Art Methods

The performance of the proposed model was assessed against state-of-the-art noise detection and correction techniques. Table 6 presents the SOC estimation results using RF regression as the evaluation model. RF regression was chosen due to its simplicity, robustness, capability to handle high-dimensional data, and its widespread use as a benchmark in SOC estimation studies.

Table 6. Comparison of noise detection and correction methods for SOC estimation using Random Forest Regression using Landsat 8 data. Performance is evaluated using R^2, RMSE, and RPD metrics. The last three columns (ΔR^2, ΔRMSE, and ΔRPD) show the percentage difference relative to the proposed method, with positive values indicating performance gaps.

Method	Noise-Detection Approach	Noise-Correction Approach	R^2 ↑	RMSE ↓	RPD ↑	ΔR^2 (%)	ΔRMSE (%)	ΔRPD (%)
Threshold-Based Exclusion	NDVI [19]	None	0.55	6.80	1.50	21.82	14.71	22.67
Statistical Detection	Z-Score [26]	Kriging [27,36]	0.61	6.68	1.67	9.84	13.17	10.18
Statistical Detection	MAD [28,29]	Kriging [27,36]	0.62	5.88	1.62	8.06	1.36	13.58
ML Detection	LOF [30,31]	Robust PCA [34,35]	0.59	6.51	1.68	13.56	10.91	9.52
Hybrid ML	OC-SVM [32,33]	Kriging [27,36]	0.66	6.07	1.75	1.52	4.45	5.14
DL Detection	VAEs [37,38]	cVAEs [39]	0.54	7.28	1.53	24.07	20.33	20.26
Proposed Method	Transformer-guided Isolation Forest	cGAN	**0.67**	**5.80**	**1.84**	–	–	–

- Threshold-Based Methods: NDVI-based exclusion is a widely used filtering technique in SOC estimation tasks [19]. These methods rely on predefined thresholds to reduce vegetation-related noise. While effective in datasets dominated by vegetation noise, they struggle with more complex noise sources, such as atmospheric effects or sensor anomalies, leading to lower accuracy in mixed-noise conditions.
- Statistical Detection and Reconstruction: Statistical techniques such as Z-Score [26,27] and MAD [28,36] identify noise based on deviations from statistical norms. When combined with kriging for reconstruction [27,36], these methods improve SOC estimation by leveraging spatial correlations. However, their reliance on assumptions like normal data distribution makes them less effective for heterogeneous noise patterns.
- Machine Learning-Based Methods: ML-based approaches such as LOF [30,31] and OC-SVM [32,33] can adapt to complex noise structures by modeling local and global anomalies. However, their performance depends on proper parameter tuning and dataset characteristics. While robust PCA [34,35] and kriging improve reconstruction, these methods remain computationally intensive and less effective for highly variable noise distributions.
- Deep Learning Approaches: DL methods, including VAEs [37,38] and cVAEs [39], leverage neural networks to model and correct noise. However, they require large, high-quality training datasets to generalize effectively. In real-world remote sensing applications, their tendency to overfit and struggle with highly variable noise distributions limits their effectiveness.
- Proposed Method: Our approach integrates a Transformer-based Isolation Forest for noise detection and a cGAN for noise correction. Unlike traditional threshold-based exclusion methods, it preserves critical spectral information by reconstructing noisy samples rather than discarding them. By leveraging deep learning for both feature extraction and anomaly detection, combined with generative models for reconstruction,

this framework ensures adaptability across different remote sensing environments, significantly improving SOC estimation accuracy in complex datasets.

The comparative study emphasizes the strengths and weaknesses of existing noise detection and correction methods. Traditional statistical techniques, or ML and DL methods, provide fundamental solutions. However, these approaches often struggle with complex satellite data patterns and effectively find heterogeneous noise. The proposed hybrid model effectively bridges the gap and provides a solution for handling noise in the dataset to achieve improved SOC estimation. Unlike traditional methods that rely solely on statistical thresholds or black-box deep learning models, the proposed approach balances interpretability and adaptability, allowing it to generalize across different datasets while maintaining computational efficiency. This ensures its practical applicability in large-scale SOC mapping using remote sensing data.

The comparative study emphasizes the strengths and weaknesses of existing noise detection and correction methods. Traditional statistical techniques, as well as ML and DL-based models, offer foundational solutions but often struggle with complex satellite reflectance patterns and heterogeneous noise. Among these, the Hybrid ML approach using OC-SVM and Kriging was previously the best-performing method. However, the proposed model outperforms it with a 5.14% improvement in RPD and a 4.45% reduction in RMSE—both crucial indicators of estimation accuracy. Although the increase in R^2 is relatively modest at 1.52%, this metric is sensitive to outliers, which are expected in noisy satellite datasets. Despite possible remaining residual noise, the consistent improvements across multiple evaluation metrics underscore the effectiveness and robustness of the proposed framework. These results confirm the superiority of our method in handling noise and improving SOC estimation for large-scale remote sensing applications.

5.2. Impact of Noise Ratio on Model Performance

The original dataset consists of 510 samples, and a synthetic dataset was systematically constructed to analyze the impact of different noise ratios. According to the Isolation Forest algorithm, 51 samples (10% of the dataset) were identified as noisy. These noisy samples were then used to generate datasets with varying noise ratios (20%, 25%, 33%, 50%, and 100%).

In the case of 100% noise, the dataset contained only the 51 noisy samples. For the 50% noise scenario, these 51 noisy samples were merged with 51 non-noisy samples. This process was repeated iteratively for the 33%, 25%, and 20% noise ratios by incrementally adding the same number of non-noisy samples. To maintain statistical consistency, the augmented dataset followed a Random Sampling with Constraints approach [67], ensuring unbiased sample selection and real-world representativeness.

The results in Table 7 highlight the impact of noise ratios on the performance of ML and DL models for SOC estimation. In this table, Baseline refers to models trained on NDVI-filtered samples without any additional noise correction, representing the traditional exclusion-based approach widely used in the literature. As noise levels increased, SOC estimation accuracy declined under the baseline condition due to the absence of noise correction, with models trained on datasets containing 100% noisy samples exhibiting the poorest performance.

Table 7. Regression results (R^2, *RMSE*, *RPD*) of ML and DL models in estimating soil organic carbon from Landsat 8 data at varying noise ratios. Both Baseline and Proposed use NDVI-filtered samples; Baseline applies no further correction, while Proposed includes the developed noise detection and correction framework.

Noise Level	Input	LR	SVR	ANN	DT	KNN	1dCNN	GB	RF	CBR
100% Noise	Baseline	−0.51	−0.12	−0.25	−0.47	−0.12	−0.23	0.16	0.07	0.31
		17.28	17.01	16.89	18.32	15.62	16.79	14.01	15.09	13.20
		0.90	0.94	0.91	0.93	1.02	0.92	1.15	1.05	1.29
	Proposed	0.61	0.48	0.59	0.81	0.75	0.50	**0.86**	0.72	0.86
		8.94	12.17	9.41	6.16	8.21	10.36	**5.20**	6.75	5.63
		1.73	1.49	1.63	3.03	2.01	1.51	**3.01**	2.47	3.00
50% Noise	Baseline	−0.51	0.08	−0.04	0.33	0.38	0.00	0.40	0.36	0.44
		16.88	16.05	15.83	12.31	12.33	15.40	11.91	12.46	11.72
		0.97	1.04	1.03	1.39	1.38	1.08	1.41	1.33	1.41
	Proposed	0.39	0.32	0.36	−0.10	0.48	0.42	0.51	0.34	**0.67**
		11.93	13.95	11.19	13.49	10.67	11.39	10.72	11.04	**8.93**
		1.36	1.23	1.48	1.40	1.61	1.45	1.54	1.53	**1.91**
33% Noise	Baseline	0.12	0.20	0.22	0.17	0.36	0.21	0.48	0.28	0.58
		11.45	11.45	10.86	10.35	9.69	10.59	8.15	9.40	7.95
		1.10	1.12	1.18	1.24	1.31	1.20	1.63	1.38	1.63
	Proposed	0.42	0.36	0.42	0.46	0.44	0.44	0.58	0.55	**0.63**
		9.64	10.30	9.43	9.25	9.20	9.13	7.93	8.19	**7.72**
		1.33	1.27	1.34	1.39	1.40	1.39	1.60	1.54	**1.66**
25% Noise	Baseline	0.12	0.28	0.21	0.00	0.33	0.30	0.42	0.49	0.57
		11.38	10.78	10.81	10.60	9.26	10.48	8.57	8.12	7.55
		1.07	1.18	1.13	1.20	1.34	1.21	1.42	1.51	1.68
	Proposed	0.47	0.36	0.53	0.29	0.52	0.43	0.56	0.50	**0.64**
		9.68	10.72	9.10	11.01	9.03	9.82	8.39	8.90	**7.66**
		1.38	1.26	1.48	1.26	1.52	1.38	1.67	1.60	**1.83**
20% Noise	Baseline	0.12	0.26	0.24	0.20	0.20	0.13	0.53	0.36	0.50
		10.86	10.47	10.03	9.41	9.48	10.70	7.71	8.27	7.94
		1.24	1.94	1.22	1.33	1.35	1.19	1.59	1.58	1.57
	Proposed	0.38	0.36	0.45	0.33	0.38	0.50	0.58	**0.58**	0.55
		9.39	9.90	8.91	8.72	9.06	8.49	7.50	**7.30**	7.78
		1.31	1.26	1.36	1.45	1.36	1.43	1.62	**1.69**	1.61

All Landsat 8 samples were pre-filtered using NDVI thresholding ($NDVI < 0.3$). Bold indicates the best performance within each noise scenario.

On the other hand, the proposed method, which employs a cGAN for noise correction, consistently outperformed models trained on noisy datasets across all noise levels. Even in extreme cases (e.g., 100% noise), the cGAN-based approach demonstrated substantial improvements over baseline models, highlighting its ability to recover meaningful spectral features from heavily noisy data. These trends are visually illustrated in Figure 7, where the x-axis represents the percentage of noise and the grouped bars show the comparison of R^2, RMSE, and RPD between the baseline and proposed methods. RF was selected for this figure because it exhibited consistently strong performance in previous experiments. The proposed approach maintains significantly higher accuracy and robustness across all noise levels, especially under severe noise conditions. The method also maintained comparable SOC estimation accuracy at noise levels as low as 20–50%, demonstrating its robustness when noise-free data are limited or difficult to obtain.

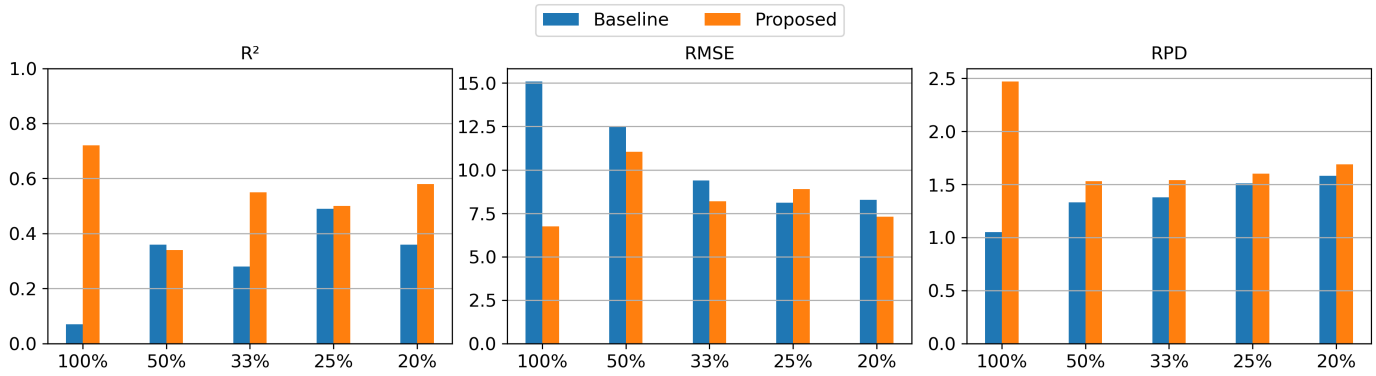


Figure 7. Random Forest performance under varying noise levels. The x-axis represents the percentage of noise introduced into the input reflectance data. The results demonstrate the effect of different noise ratios on model accuracy (R^2), estimation error (RMSE), and robustness (RPD) for both baseline and proposed methods.

By correcting noisy samples rather than discarding them, the proposed framework maximizes dataset utility, enhances model generalizability, and provides a scalable solution for real-world SOC estimation challenges.

5.3. Evaluation Under High Vegetation Cover Using Sentinel-2

This subsection evaluates the performance of various ML and DL models for SOC estimation using the S2 dataset, with a particular focus on testing the proposed method under high vegetation cover conditions. This analysis demonstrates the applicability of the framework beyond bare soil scenarios and assesses its robustness in more complex, vegetation-influenced environments.

Unlike the L8 dataset, which focused on bare soil samples, the S2 dataset includes pixels with substantial vegetation cover (NDVI up to 0.93). This design enables evaluation of the model's capability to handle vegetation-contaminated reflectance and estimate SOC in mixed soil–vegetation conditions, which more closely represent real-world agricultural landscapes.

Samples with mixed soil and vegetation conditions ($0 < \text{NDVI} < 0.93$) were purposefully included to challenge the model with vegetation-induced spectral noise, which represents one of the most difficult real-world scenarios for SOC estimation. Table 8 provides a summary of the dataset, highlighting its high coefficient of variation (CV = 95.28%), which underscores the substantial variability arising from differing soil types and vegetation conditions.

Table 8. Descriptive statistical parameters for the SOC dataset investigated in this study using Sentinel-2 data.

Soil Type	NDVI Range	S. No.	Min	Max	Mean	Median	Std.	CV(%)
Mixed	$0 < \text{NDVI} < 0.93$	485	2.3	172.3	17.12	13.80	16.37	95.28

Table 9 presents the regression results across all models. Models trained on raw reflectance data showed poor performance due to the high level of spectral noise introduced by vegetation and other distortions. The best existing approach (Table 6) in the literature that uses OC-SVM [32,33] for noise detection and kriging [27,36] for reconstruction provided only moderate improvement. This is primarily due to OC-SVM's limited ability to capture complex nonlinear spectral patterns and Kriging's reliance on spatial continuity, which is often disrupted in noisy reflectance data.

Table 9. Regression results (R^2 ↑, $RMSE$ ↓, RPD ↑) of ML and DL models in estimating soil organic carbon using Sentinel-2 dataset.

Input	LR	SVR	ANN	DT	KNN	1D CNN	GB	RF	CBR
Raw Data	0.04 15.14 1.03	0.02 15.66 1.01	0.07 15.01 1.04	−1.21 20.49 0.78	−0.02 15.47 1.01	0.02 15.21 1.03	0.00 15.37 1.02	−0.07 15.79 0.99	−0.02 15.62 1.00
OC-SVM & Kriging	0.10 11.43 1.09	0.03 15.64 1.02	0.05 15.04 1.05	0.04 19.49 0.98	0.04 15.23 1.00	0.14 15.36 1.13	0.12 14.79 1.11	0.13 14.68 1.03	0.15 14.78 1.10
Proposed	0.20 11.09 1.22	0.12 15.94 1.11	0.21 15.76 1.13	0.21 19.68 1.12	0.26 13.50 1.20	**0.39** **9.51** **1.67**	0.32 12.47 1.26	0.37 11.95 1.34	0.37 12.11 1.37

Bold = best performing model per metric. OC-SVM [32,33] + Kriging [27,36] results were reproduced in this study based on the original algorithm descriptions. The proposed approach was benchmarked against this method, which showed the best performance among existing models in Table 6.

In contrast, the proposed framework substantially outperformed both baselines by effectively suppressing noise and enhancing SOC estimation accuracy. The best performance was achieved by the 1D CNN model, with an R^2 of 0.39, RMSE of 9.51, and RPD of 1.67 when integrated with the proposed approach.

These findings demonstrate that the proposed approach is not restricted to bare soil conditions but can also be applied to vegetated surfaces, effectively mitigating vegetation-induced spectral distortions and maintaining reliable SOC estimation. This highlights its potential for broader applications in soil monitoring and carbon estimation, particularly in regions with substantial vegetation cover.

5.4. Practical Implications

The proposed methodology possesses considerable practical importance for SOC estimation based on remote sensing data. This approach can improve SOC estimation by using a Transformer-guided Isolation Forest to avoid noise influences, while cGAN can restore useful information from noisy data. Its ability to outperform traditional and state-of-the-art methods in terms of R^2, RMSE, and RPD demonstrate its suitability for large-scale applications, such as regional or national SOC monitoring programs. In addition, the use of DL methods such as cGAN provides a scalable solution capable of managing high-dimensional and intricate satellite datasets, alleviating the reliance on time-consuming data cleaning and preprocessing efforts. This functional benefit makes it an efficient tool for policymakers and agricultural interest groups who wish to promote data-driven soil management policy.

Furthermore, the adaptability of the proposed method extends to various geographic regions, soil types, and vegetation covers, making it highly applicable across diverse environmental conditions. While the S2 validation supports its generalization across different remote sensing platforms, additional validation using field-collected datasets from heterogeneous agro-ecological zones is necessary to fully assess its robustness. Such validation efforts will help ensure its reliability for large-scale soil monitoring programs, providing valuable insights for precision agriculture and environmental sustainability.

5.5. Limitations and Future Work

The proposed approach demonstrates robustness, even in heavily noisy datasets. However, certain limitations remain. One key challenge is the computational cost of the cGAN-based reconstruction, which could be further optimized to enable scalability for large datasets and potential real-time applications. Furthermore, reliance on a specific feature

extraction technique, the transformer-guided isolation forest, may limit its adaptability to datasets with varying spectral properties and feature distributions. Exploring alternative DL architectures or hybrid approaches that integrate domain-specific spectral characteristics could enhance the generalizability of the method.

While this study focuses on SOC estimation, extending the framework to estimate other soil properties or environmental variables, such as soil texture or nutrient concentrations, would broaden its applicability. Future research could explore incorporating domain knowledge, such as soil classifications or vegetation indices, directly into the reconstruction process. Integrating physics-based constraints or multi-source data fusion (e.g., combining optical and thermal data) may further improve model accuracy and reliability.

Finally, although the method has been validated using artificially synthesized noise ratios, its effectiveness on field-collected datasets with real-world noise variations remains to be tested. Future studies should assess the approach on diverse agro-ecological zones and varying soil conditions to confirm its robustness in practical applications. Additionally, evaluating hybrid models that integrate cGAN with statistical or physics-based techniques could enhance the representation of soil reflectance dynamics, making the framework more applicable to remote sensing in agriculture and environmental monitoring under broader conditions. Furthermore, since SOC is influenced by various environmental and anthropogenic factors such as climate, terrain, and land use practices, future extensions of this framework could explore the integration of these auxiliary variables to improve the spatial generalizability and ecological interpretability of SOC estimations.

6. Conclusions

This study presents a robust methodology for estimating SOC from satellite reflectance data, integrating Transformer-based Isolation Forest for noise detection and cGAN for correcting noisy samples. By overcoming the limitations of traditional exclusion-based noise-filtering methods, this approach retains and reconstructs valuable information, significantly improving data utility and SOC estimation accuracy. Validation across L8 and S2 datasets demonstrates its adaptability to varying noise patterns and remote sensing platforms, offering a scalable and effective solution for addressing noisy data in environmental and agricultural applications. The ability of this framework to generalize across different datasets highlights its potential for improving SOC estimation in remote sensing-based soil monitoring. A preliminary version of this work is available as a preprint [68].

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

Funding: This research was funded by Charles Sturt University, Australia, under a PhD scholarship. The funding was provided through Fund A541—School Org 2020, Program Code 25769.

Data Availability Statement: The satellite reflectance datasets used in this study, including L8 and S2 observations paired with corresponding SOC measurements, are publicly available via the project's open-access GitHub repository. The repository also contains the full implementation of the proposed noise detection and correction framework. The proposed framework and datasets can be accessed at: https://github.com/DristiDatta/Transformer_Guided_Noise_Detection [69] (accessed on 22 August 2025).

Acknowledgments: This work has been supported by the Cooperative Research Centre for High Performance Soils whose activities are funded by the Australian Government's Cooperative Research Centre Program and Charles Sturt University, Australia.

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:

1D-CNN	One-dimensional convolutional neural network
ANN	Artificial neural network
B	Band
BI	Brightness index
CBR	CatBoost regressor
cGAN	Conditional generative adversarial network
CI	Color index
cVAE	Conditional variational autoencoder
DL	Deep learning
DSI	Dry soil index
DT	Decision tree
EC	Electrical conductivity
ESDAC	European Soil Data Centre
EVI	Enhanced vegetation index
FLAASH	Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes
GB	Gradient boosting
GNDVI	Green normalized difference vegetation index
HSI	Hyperspectral imaging
K	Potassium
KNN	k-Nearest neighbors
L8	Landsat 8
LOF	Local outlier factor
LR	Linear regression
LUCAS	Land Use/Cover Area frame Survey
MAD	Median absolute deviation
ML	Machine learning
MSAVI	Modified soil-adjusted vegetation index
N	Nitrogen
NDVI	Normalized difference vegetation index
OC-SVM	One-class support vector machine
P	Phosphorus
PCA	Principal component analysis
RF	Random forest
RVI	Ratio vegetation index
S2	Sentinel-2
SAVI	Soil-adjusted vegetation index
SC	Scenario
SI	Salinity index
SOC	Soil organic carbon
SVR	Support vector regression
TCT	Tasseled Cap transformation
VAE	Variational autoencoder
VIs	Vegetation indices

References

1. Fageria, N. Role of soil organic matter in maintaining sustainability of cropping systems. *Commun. Soil Sci. Plant Anal.* **2012**, *43*, 2063–2113. [CrossRef]
2. Bhattacharya, S.S.; Kim, K.H.; Das, S.; Uchimiya, M.; Jeon, B.H.; Kwon, E.; Szulejko, J.E. A review on the role of organic inputs in maintaining the soil carbon pool of the terrestrial ecosystem. *J. Environ. Manag.* **2016**, *167*, 214–227. [CrossRef]
3. Weil, R.R.; Islam, K.R.; Stine, M.A.; Gruver, J.B.; Samson-Liebig, S.E. Estimating active carbon for soil quality assessment: A simplified method for laboratory and field use. *Am. J. Altern. Agric.* **2003**, *18*, 3–17. [CrossRef]
4. Loria, N.; Lal, R.; Chandra, R. Handheld In Situ Methods for Soil Organic Carbon Assessment. *Sustainability* **2024**, *16*, 5592. [CrossRef]
5. Liu, S.; Chen, J.; Guo, L.; Wang, J.; Zhou, Z.; Luo, J.; Yang, R. Prediction of soil organic carbon in soil profiles based on visible–near-infrared hyperspectral imaging spectroscopy. *Soil Tillage Res.* **2023**, *232*, 105736. [CrossRef]
6. Li, Y.; Chang, C.; Wang, Z.; Zhao, G. Remote sensing prediction and characteristic analysis of cultivated land salinization in different seasons and multiple soil layers in the coastal area. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *111*, 102838. [CrossRef]
7. Ge, X.; Ding, J.; Jin, X.; Wang, J.; Chen, X.; Li, X.; Liu, J.; Xie, B. Estimating agricultural soil moisture content through UAV-based hyperspectral images in the arid region. *Remote Sens.* **2021**, *13*, 1562. [CrossRef]
8. Datta, D.; Paul, M.; Murshed, M.; Teng, S.W.; Schmidtke, L. Soil Moisture, Organic Carbon, and Nitrogen Content Prediction with Hyperspectral Data Using Regression Models. *Sensors* **2022**, *22*, 7998. [CrossRef] [PubMed]
9. Sargeant, J.; Teng, S.W.; Murshed, M.; Paul, M.; Brennan, D. Estimating Soil Organic Carbon from Multispectral Images Using Physics-Informed Neural Networks. In Proceedings of the Asian Conference on Computer Vision, Hanoi, Vietnam, 8–12 December 2024; pp. 2632–2649.
10. Rahman, M.; Teng, S.W.; Murshed, M.; Paul, M.; Brennan, D. Deep Learning-based Adaptive Downsampling of Hyperspectral Bands for Soil Organic Carbon Estimation. *IEEE Access* **2025**, *13*, 95392–95409. [CrossRef]
11. Rahman, M.; Teng, S.W.; Murshed, M.; Paul, M.; Brennan, D. Addressing Limitations of Common Methods in Attention-Based Hyperspectral Band Selection Algorithms. In Proceedings of the 2024 IEEE International Conference on Digital Image Computing: Techniques and Applications (DICTA), Perth, Australia, 27–29 November 2024; pp. 640–647.
12. Rahman, M.; Teng, S.W.; Murshed, M.; Paul, M.; Brennan, D. BSDR: A data-efficient deep learning-based hyperspectral band selection algorithm using discrete relaxation. *Sensors* **2024**, *24*, 7771. [CrossRef]
13. Datta, D.; Paul, M.; Murshed, M.; Teng, S.W.; Schmidtke, L. Comparative Analysis of Machine and Deep Learning Models for Soil Properties Prediction from Hyperspectral Visual Band. *Environments* **2023**, *10*, 77. [CrossRef]
14. Stiglitz, R.; Mikhailova, E.; Post, C.; Schlautman, M.; Sharp, J. Using an inexpensive color sensor for rapid assessment of soil organic carbon. *Geoderma* **2017**, *286*, 98–103. [CrossRef]
15. Nodi, S.S.; Paul, M.; Robinson, N.; Wang, L.; Rehman, S.U. Determination of Munsell Soil Colour Using Smartphones. *Sensors* **2023**, *23*, 3181. [CrossRef] [PubMed]
16. Nodi, S.S.; Paul, M.; Robinson, N.; Wang, L.; Rehman, S.U.; Kabir, M.A. Munsell soil colour prediction from the soil and soil colour book using patching method and deep learning techniques. *Sensors* **2025**, *25*, 287. [CrossRef] [PubMed]
17. Demattê, J.A.; Poppiel, R.R.; Novais, J.J.M.; Rosin, N.A.; Minasny, B.; Savin, I.Y.; Grunwald, S.; Chen, S.; Hong, Y.; Huang, J.; et al. Frontiers in earth observation for global soil properties assessment linked to environmental and socio-economic factors. *Innovation* **2025**, *6*, 100985. [CrossRef]
18. Pande, C.B.; Kadam, S.A.; Jayaraman, R.; Gorantiwar, S.; Shinde, M. Prediction of soil chemical properties using multispectral satellite images and wavelet transforms methods. *J. Saudi Soc. Agric. Sci.* **2022**, *21*, 21–28. [CrossRef]
19. Datta, D.; Paul, M.; Murshed, M.; Teng, S.W.; Schmidtke, L.M. Novel Dry Soil and Vegetation Indices to Predict Soil Contents from Landsat 8 Satellite Data. In Proceedings of the 2023 IEEE International Conference on Digital Image Computing: Techniques and Applications (DICTA), Port Macquarie, Australia, 28 November–1 December 2023; pp. 152–159.
20. Yuzugullu, O.; Fajraoui, N.; Don, A.; Liebisch, F. Satellite-based soil organic carbon mapping on European soils using available datasets and support sampling. *Sci. Remote Sens.* **2024**, *9*, 100118. [CrossRef]
21. Pettorelli, N. *The Normalized Difference Vegetation Index*; Oxford University Press: Oxford, UK, 2013.
22. Xue, J.; Su, B. Significant remote sensing vegetation indices: A review of developments and applications. *J. Sens.* **2017**, *2017*, 1353691. [CrossRef]
23. Liao, Z.; He, B.; Quan, X. Modified enhanced vegetation index for reducing topographic effects. *J. Appl. Remote Sens.* **2015**, *9*, 096068. [CrossRef]
24. Muzhoffar, D.A.F.; Sakuno, Y.; Taniguchi, N.; Hamada, K.; Shimabukuro, H.; Hori, M. Automatic Detection of Floating Macroalgae via Adaptive Thresholding Using Sentinel-2 Satellite Data with 10 m Spatial Resolution. *Remote Sens.* **2023**, *15*, 2039. [CrossRef]
25. Lee, J.K.; Acharya, T.D.; Lee, D.H. Exploring Land Cover Classification Accuracy of Landsat 8 Image Using Spectral Index Layer Stacking in Hilly Region of South Korea. *Sens. Mater.* **2018**, *30*, 2927–2941. [CrossRef]
26. Shiffler, R.E. Maximum Z scores and outliers. *Am. Stat.* **1988**, *42*, 79–80. [CrossRef]

27. Ma, Y.; Ma, Y. Geostatistical estimation methods: Kriging. In *Quantitative Geosciences: Data Analytics, Geostatistics, Reservoir Characterization and Modeling*; Springer: Berlin/Heidelberg, 2019; pp. 373–401.
28. Voloh, B.; Watson, M.R.; König, S.; Womelsdorf, T. MAD saccade: Statistically robust saccade threshold estimation via the median absolute deviation. *J. Eye Mov. Res.* **2020**, *12*, 1–12. [CrossRef] [PubMed]
29. Li, Y.; Li, Z.; Wei, K.; Xiong, W.; Yu, J.; Qi, B. Noise estimation for image sensor based on local entropy and median absolute deviation. *Sensors* **2019**, *19*, 339. [CrossRef]
30. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16–18 May 2000; pp. 93–104.
31. Campos, G.O.; Zimek, A.; Sander, J.; Campello, R.J.; Micenková, B.; Schubert, E.; Assent, I.; Houle, M.E. On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Min. Knowl. Discov.* **2016**, *30*, 891–927. [CrossRef]
32. Ananias, P.H.M.; Negri, R.G. Anomalous behaviour detection using one-class support vector machine and remote sensing images: A case study of algal bloom occurrence in inland waters. *Int. J. Digit. Earth* **2021**, *14*, 921–942. [CrossRef]
33. Alam, S.; Sonbhadra, S.K.; Agarwal, S.; Nagabhushan, P. One-class support vector classifiers: A survey. *Knowl.-Based Syst.* **2020**, *196*, 105754. [CrossRef]
34. Candès, E.J.; Li, X.; Ma, Y.; Wright, J. Robust principal component analysis? *J. ACM* **2011**, *58*, 1–37. [CrossRef]
35. Zhou, T.; Tao, D. Godec: Randomized low-rank & sparse matrix decomposition in noisy case. In Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, WA, USA, 28 June 28–2 July 2011.
36. Lin, Q.; Li, C. Kriging based sequence interpolation and probability distribution correction for gaussian wind field data reconstruction. *J. Wind. Eng. Ind. Aerodyn.* **2020**, *205*, 104340. [CrossRef]
37. Nowroozilarki, Z.; Mortazavi, B.J.; Jafari, R. Variational autoencoders for biomedical signal morphology clustering and noise detection. *IEEE J. Biomed. Health Inform.* **2023**, *28*, 169–180. [CrossRef]
38. Sadeghi, M.; Alameda-Pineda, X. Switching variational auto-encoders for noise-agnostic audio-visual speech enhancement. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, Canada, 6–11 June 2021; pp. 6663–6667.
39. Zhang, C.; Barbano, R.; Jin, B. Conditional variational autoencoder for learned image reconstruction. *Computation* **2021**, *9*, 114. [CrossRef]
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
41. Soydaner, D. Attention mechanism in neural networks: Where it comes and where it goes. *Neural Comput. Appl.* **2022**, *34*, 13371–13385. [CrossRef]
42. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
43. Xu, H.; Pang, G.; Wang, Y.; Wang, Y. Deep isolation forest for anomaly detection. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 12591–12604. [CrossRef]
44. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422.
45. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data (TKDD)* **2012**, *6*, 1–39. [CrossRef]
46. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784. [CrossRef]
47. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
48. Orgiazzi, A.; Ballabio, C.; Panagos, P.; Jones, A.; Fernández-Ugalde, O. LUCAS Soil, the largest expandable soil dataset for Europe: A review. *Eur. J. Soil Sci.* **2018**, *69*, 140–153. [CrossRef]
49. Loveland, T.R.; Irons, J.R. Landsat 8: The plans, the reality, and the legacy. *Remote Sens. Environ.* **2016**, *185*, 1–6. [CrossRef]
50. Barsi, J.A.; Schott, J.R.; Hook, S.J.; Raqueno, N.G.; Markham, B.L.; Radocinski, R.G. Landsat-8 thermal infrared sensor (TIRS) vicarious radiometric calibration. *Remote Sens.* **2014**, *6*, 11607–11626. [CrossRef]
51. Thorne, K.; Markharn, B.; Barker, J.; Slater, P.; Biggar, S. Radiometric calibration of Landsat. *Photogramm. Eng. Remote Sens.* **1997**, *63*, 853–858.
52. Gao, B.C.; Montes, M.J.; Davis, C.O.; Goetz, A.F. Atmospheric correction algorithms for hyperspectral remote sensing data of land and ocean. *Remote Sens. Environ.* **2009**, *113*, S17–S24.
53. Maurer, T. How to pan-sharpen images using the gram-schmidt pan-sharpen method–A recipe. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2013**, *40*, 239–244. [CrossRef]

54. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sens. Environ.* **1996**, *58*, 289–298.
55. Qi, J.; Chehbouni, A.; Huete, A.R.; Kerr, Y.H.; Sorooshian, S. A modified soil adjusted vegetation index. *Remote Sens. Environ.* **1994**, *48*, 119–126. [CrossRef]
56. Dehni, A.; Lounis, M. Remote sensing techniques for salt affected soil mapping: Application to the Oran region of Algeria. *Procedia Eng.* **2012**, *33*, 188–198. [CrossRef]
57. Tucker, C.J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **1979**, *8*, 127–150. [CrossRef]
58. Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring vegetation systems in the Great Plains with ERTS. *NASA Spec. Publ.* **1974**, *351*, 309.
59. Matsushita, B.; Yang, W.; Chen, J.; Onda, Y.; Qiu, G. Sensitivity of the enhanced vegetation index (EVI) and normalized difference vegetation index (NDVI) to topographic effects: A case study in high-density cypress forest. *Sensors* **2007**, *7*, 2636–2651. [CrossRef] [PubMed]
60. Rondeaux, G.; Steven, M.; Baret, F. Optimization of soil-adjusted vegetation indices. *Remote Sens. Environ.* **1996**, *55*, 95–107. [CrossRef]
61. Baig, M.H.A.; Zhang, L.; Shuai, T.; Tong, Q. Derivation of a tasselled cap transformation based on Landsat 8 at-satellite reflectance. *Remote Sens. Lett.* **2014**, *5*, 423–431. [CrossRef]
62. Mulla, D.; McBratney, A.B. *Soil Spatial Variability*; Soil physics companion; CRC Press: Boca Raton, FL, USA, 2001.
63. Fearn, T. Assessing calibrations: Sep, rpd, rer and r 2. *NIR News* **2002**, *13*, 12–13. [CrossRef]
64. Chen, Q.; Wang, Y.; Zhu, X. Soil organic carbon estimation using remote sensing data-driven machine learning. *PeerJ* **2024**, *12*, e17836. [CrossRef] [PubMed]
65. Datta, D.; Paul, M.; Murshed, M.; Teng, S.W.; Schmidtke, L.M. Unveiling Soil-Vegetation Interactions: Reflection Relationships and an Attention-Based Deep Learning Approach for Carbon Estimation. In Proceedings of the 2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Niagara Falls, ON, Canada, 29 August 2024; pp. 1–6.
66. John, K.; Abraham Isong, I.; Michael Kebonye, N.; Okon Ayito, E.; Chapman Agyeman, P.; Marcus Afu, S. Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. *Land* **2020**, *9*, 487. [CrossRef]
67. Luedtke, J.; Ahmed, S. A sample approximation approach for optimization with probabilistic constraints. *SIAM J. Optim.* **2008**, *19*, 674–699. [CrossRef]
68. Paul, M.; Datta, D.; Murshed, M.; Teng, S.W.; Schmidtke, L.M. Transformer-Guided Noise Detection and Correction in Remote Sensing Data for Enhanced Soil Organic Carbon Estimation. *SSRN* 2025, *in press*.
69. Datta, D. Transformer-Guided Noise Detection and Correction Framework. 2025. Available online: https://github.com/DristiDatta/Transformer_Guided_Noise_Detection (accessed on 22 August 2025).

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.