European Journal of Soil Science

WILEY

European Journal of Soil Science

RESEARCH ARTICLE OPEN ACCESS

Towards Explainable AI: Interpreting Soil Organic Carbon Prediction Models Using a Learning-Based Explanation Method

Nafiseh Kakhani[1,2] | Ruhollah Taghizadeh-Mehrjardi[1,2,3,4] | Davoud Omarzadeh[5] | Masahiro Ryo[6,7] | Uta Heiden[8] | Thomas Scholten[1,2,3]

[1]Department of Geosciences, Soil Science and Geomorphology, University of Tübingen, Tübingen, Germany | [2]CRC 1070 RessourceCultures, University of Tübingen, Tübingen, Germany | [3]DFG Cluster of Excellence "Machine Learning", University of Tübingen, Tübingen, Germany | [4]Faculty of Agriculture & Natural Resources, Ardakan University, Ardakan, Iran | [5]Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Barcelona, Spain | [6]Leibniz Centre for Agricultural Landscape Research (ZALF), Müncheberg, Germany | [7]Brandenburg University of Technology Cottbus–Senftenberg, Cottbus, Germany | [8]German Aerospace Center, The Remote Sensing Technology Institute, Department of Photogrammetry and Image Analysis, Wessling, Germany

Correspondence: Nafiseh Kakhani (nafiseh.kakhani@uni-tuebingen.de)

Received: 1 August 2024 | **Revised:** 28 November 2024 | **Accepted:** 31 January 2025

Funding: This work was supported by the Deutsche Forschungsgemeinschaft.

Keywords: explainable AI | Germany | Google Earth Engine | HLS product | remote sensing | soil organic carbon

ABSTRACT

An understanding of the key factors and processes influencing the variability of soil organic carbon (SOC) is essential for the development of effective policies aimed at enhancing carbon storage in soils to mitigate climate change. In recent years, complex computational approaches from the field of machine learning (ML) have been developed for modelling and mapping SOC in various ecosystems and over large areas. However, in order to understand the processes that account for SOC variability from ML models and to serve as a basis for new scientific discoveries, the predictions made by these data-driven models must be accurately explained and interpreted. In this research, we introduce a novel explanation approach applicable to any ML model and investigate the significance of environmental features to explain SOC variability across Germany. The methodology employed in this study involves training multiple ML models using SOC content measurements from the LUCAS dataset and incorporating environmental features derived from Google Earth Engine (GEE) as explanatory variables. Thereafter, an explanation model is applied to elucidate what the ML models have learned about the relationship between environmental features and SOC content in a supervised manner. In our approach, a post hoc model is trained to estimate the contribution of specific inputs to the outputs of the trained ML models. The results of this study indicate that different classes of ML models rely on interpretable but distinct environmental features to explain SOC variability. Decision tree-based models, such as random forest (RF) and gradient boosting, highlight the importance of topographic features. Conversely, soil chemical information, particularly pH, is crucial for the performance of neural networks and linear regression models. Therefore, interpreting data-driven studies requires a carefully structured approach, guided by expert knowledge and a deep understanding of the models being analysed.

Summary

- The integration of machine learning in environmental and soil sciences advances research but lacks transparency in decision-making processes, necessitating model explainability.
- Framed as a supervised learning task, we aim to train an explanation model to quantify input influence on ML model outputs, supplemented with expert-based environmental interpretations.
- Different ML model classes rely on distinct environmental features to explain SOC variability. Decision tree-based models (e.g., random forest, gradient boosting) highlight topographic features, while neural networks and linear regression models depend on soil chemical information.
- Interpreting data-driven studies requires a structured approach, guided by expert knowledge and a deep understanding of the models analysed.

1 | Introduction

Soil organic carbon (SOC) is critical for maintaining soil health and fertility, enhancing its structure, water retention and nutrient availability. SOC plays a significant role in mitigating global warming by sequestering atmospheric CO_2, thereby reducing greenhouse gas concentrations (Lal 2004; Zeraatpisheh et al. 2022). Additionally, SOC supports biodiversity by providing energy and habitat for microorganisms, which are essential for nutrient cycling and decomposition processes (Powlson et al. 2011). Effective management of SOC is vital for sustainable agricultural practices and long-term food security (Smith et al. 2008). In this context, modelling SOC using machine learning (ML) is critical for accurately predicting SOC dynamics and supporting more effective soil management, particularly when accounting for both spatial and temporal variations in 4D mapping (e.g., Gerke et al. 2022).

ML technologies are now extensively applied across various domains. This advancement has significantly boosted the integration of ML in scientific research. Typically, ML models are trained to achieve high accuracy, and there is a growing demand to grasp model functioning and rationalise decisions (Burkart and Huber 2021). This is crucial, since many scientists apply ML methodologies to optimise/generate scientific findings (Roscher et al. 2020). Interpretability is paramount to the scientific validity of results to derive new insights and discoveries from observed or simulated data (Samek et al. 2017). Recent advancements in this field have significantly influenced other scientific disciplines such as soil science and biogeochemistry through the availability of new tools to leverage explainable AI, providing deeper insights and understanding of complex datasets.

Many studies have applied explainable AI to understand and predict various soil and ecological processes (e.g., Wadoux and Molnar 2022). One of the most frequently used approaches is Shapley Additive Explanation (SHAP) (Lundberg and Lee 2017). For example, a study developed a Shapley value-based approach to interpret ML spectroscopic models. When applied to an RF prediction model for SOC, Shapley values provided dominant spectral contributions, increasing understanding and trust in soil spectroscopy predictions (Wadoux 2023). While SHAP values are widely used for model interpretability, they have notable limitations. They are computationally intensive, especially for large datasets and complex models (Vowels 2022). Additionally, SHAP relies on approximation methods such as KernelSHAP and TreeSHAP to make the calculations feasible, which can introduce inaccuracies (Sundararajan and Najmi 2020). Wadoux et al. (2020) explored the use of ML for digital soil mapping (DSM) and highlighted methods for interpreting deep learning models. They specifically discussed the use of feature importance metrics and partial dependence plots (PDPs) and accumulated local effects plots for model explainability. However, for models that make abrupt or highly non-linear predictions (e.g., decision trees with deep splits), PDPs may not accurately capture the relationship between features and the outcome due to their inability to reflect complex interactions in the data (Vowels 2022).

Another research effort developed a physics-guided ML approach that integrated physical parameters into ML models to improve the monitoring and prediction of soil and environmental dynamics. This method combines simulated data with various physical parameters to enhance model accuracy and reliability, leveraging scientific knowledge to better understand and manage soil processes (Chen et al. 2023). In another study, the coupling of microbial-explicit models with ML improved the simulation of SOC turnover, demonstrating the synergy between domain-specific knowledge and advanced algorithms (Xu et al. 2024). These hybrid modelling approaches align with explainable AI by incorporating known physical laws and mechanisms into ML models, ensuring that predictions are both accurate and interpretable. However, physics-based ML approaches require significant expert knowledge and are often limited to specific scenarios.

Model-specific feature importance estimations are widely used for explaining ML models. For instance, these methods were applied in France (Mulder et al. 2015) and utilised for agricultural soils in Germany (Vos et al. 2019). These methods generate statistics that assess the relative importance of different features within the model. While these indicators are both valid and valuable, they only provide a global measure of variable importance across the entire study area. They do not consider spatial variations in the interaction between environmental factors and SOC contents, which can lead to regional differences in modelling (Wadoux et al. 2022). In a recent study, activation maps and a local error-correction mechanism have been proposed for elucidating deep learning models (Tziolas et al. 2024). However, this method also requires generalisation to be applicable to models beyond deep learning. Despite these efforts, a critical research gap remains in systematically comparing feature importance measures across different ML algorithms.

Another approach to model interpretation is to treat the task of providing explanations for ML model decisions as a learning problem, training an explanation model to estimate the influence of specific inputs on another ML model's outputs (Schwab et al. 2019). One such method is CXPlain (Schwab and Karlen 2019), which uses a supervised learning process to deliver post hoc explanations—i.e., explanations generated after the model's predictions—without requiring modifications to the original model. Such an explanation model employs an objective to train a supervised ML model

to explain another ML model (Covert et al. 2021). This approach has several advantages over previously described methods. It is applicable to any ML model and data modality, as it does not require modifying or retraining the original prediction model. Additionally, it can be computationally efficient.

This method has been widely applied across various contexts such as deep learning and computer vision and is well-received in the research community for its flexibility and effectiveness. For instance, in Chen et al. (2024), it was used to explain graph neural networks (GNNs) by focusing on the most relevant graph structures influencing predictions. The approach generates counterfactual and model-level explanations while ensuring reliability by keeping the explanations aligned with the underlying data distribution. In another paper (Situ et al. 2021), the authors propose a method called Learning to Explain (L2E), which uses CXPlain as a base approach to generate explanations for black-box models. Instead of directly explaining the model output, L2E distils the explanation algorithm into a separate explainer network. This allows for the generation of more stable and faster explanations by learning the behaviour of the underlying model and applying it to new instances. Inspired by this methodology, other researchers in Chuang et al. (2023) introduced a similar L2E concept, where they trained an explanation model to mimic the behaviour of an existing explanation algorithm. They designed a framework that uses an explanation encoder to learn latent explanations through positive and negative sampling strategies based on contrastive learning. The authors in Hostallero et al. (2023) used CXPlain to make their deep learning framework interpretable. CXPlain was leveraged to help identify key genes that influence the drug response predictions made by the model. This interpretability allowed them to highlight a small set of genes whose expression levels are crucial for predicting drug sensitivity, thus aiding in the identification of biomarkers related to drug response.

In this paper, we demonstrate how the aforementioned explanation model can assist in explaining the association between a soil property (SOC) and environmental factors identified by an ML model. After preparing a comprehensive set of environmental feature inputs, we applied several popular regression models, including decision trees, neural networks (NNs) and linear regression, to generate a map of SOC content across Germany. Using multiple models ensures a robust and comprehensive analysis by allowing result comparisons and verifying reliability. We then implemented the explanation model to compute the importance score of each feature. The outcome of the explanation model, specifically the feature importance, provides insights into which environmental factors most significantly influence SOC content. We complemented our analysis by interpreting the results based on expert knowledge towards understanding the key factors and processes influencing the variability of SOC and compared them with previous studies that predict SOC content over Germany.

2 | Datasets

2.1 | Ground Reference Samples

The LUCAS Programme was initiated in 2001 as a Eurostat-managed area frame survey by the statistical office of the EU. The survey is based on the visual evaluation of agricultural policy-relevant factors. Since 2006, sampling has been conducted at the intersections of a regular 22 km grid covering the EU's territory. Eurostat, together with the European Commission's Directorates-General for the Environment and the Joint Research Centre, designed a topsoil assessment component ('LUCAS-Topsoil') within the LUCAS survey (Toth et al. 2013; Ballabio et al. 2016). This component, with a sampling depth of 0–20 cm, was created to produce the first harmonised and comparable data on soil at the European level to support policymaking (Orgiazzi et al. 2018). In our analysis, we utilised this dataset for both 2015 and 2018, which contains SOC information for all EU countries. However, we restricted our study area to Germany only to focus on the specific factors affecting SOC estimation in this region. The spatial distribution of 1686 samples throughout Germany is displayed in Figure 1. The soil samples exhibit severe skewness, with fewer samples in the upper quantiles, as shown in Figure 2. Additionally, Figure 3 shows the distribution of land cover classes at 200,000 randomly selected locations across Germany (Table 1).

2.2 | Input Features

In this study, we employed five distinct categories of environmental features: soil information, remote sensing images, vegetation, climate and topography. These features are designed to quantify SOC by providing insights into the key biotic and abiotic processes that influence SOC content, utilising the expertise and domain knowledge of field specialists (Sakhaee et al. 2022). We followed a meticulous procedure to prepare these features, which will be thoroughly discussed in the following sections. Google Earth Engine (GEE), leveraging Google's computing infrastructure and publicly accessible remote sensing datasets,

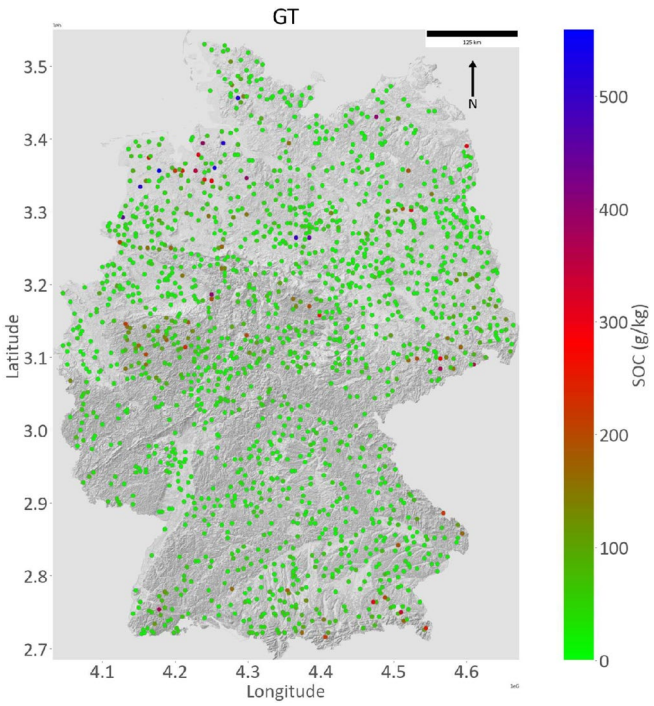

FIGURE 1 | The spatial distribution of SOC ground truth (GT) samples.

was the primary source for all the features used (Figure 4). A comprehensive list of all input features is provided in Table 2, and the final spatial resolution of the analysis is 250 m. To facilitate further exploration and adoption of our analysis, we have made it freely available at: https://github.com/nafisehkakhani/XAI-for-SOC-models.

2.2.1 | Soil Information

Eight features representing various aspects of soil properties were included: the map of clay content, as it directly correlates with SOC (Ballabio et al. 2016); the map of pH, since soil acidity affects microbial activities that drive soil organic matter turnover, thereby influencing SOC (Beugnon et al. 2023; Malik et al. 2018) and the map of water content, due to its interaction with SOC through plant productivity (Ballabio et al. 2016). Additionally, we considered soil taxonomy and texture maps, as they differentiate soil types and explain their geographical distribution across the nation (Yu et al. 2020). These maps are accessible via GEE (Hengl et al. 2017).

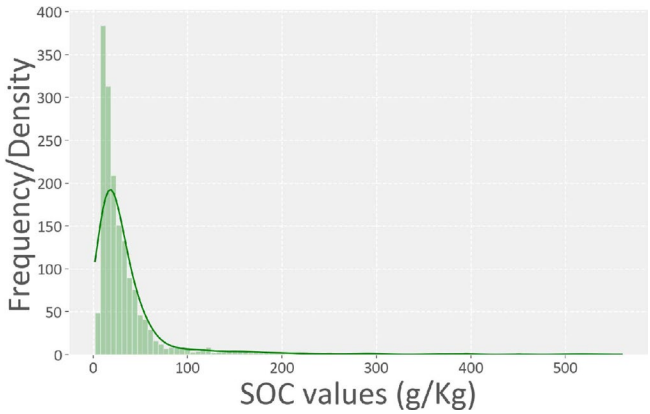

FIGURE 2 | Histogram and KDE plot of SOC values.

2.2.2 | Remote Sensing Images

Publicly available remote sensing images were used for predictions in numerous studies in soil science, following detailed preprocessing steps. For instance, in a study by Broeg et al. (2024) soil reflectance composites based on Landsat images were used for large-scale predictions of soil properties. Similarly, another study by Wang et al. (2021) carefully prepared Sentinel-2 images to predict soil organic matter. In this study, we used the Harmonised Landsat and Sentinel-2 (HLS) dataset, which combines these two prominent satellite imagery sources, for our analysis. The HLS project, a NASA initiative, aims to produce a virtual constellation of surface reflectance data from the Operational Land Imager (OLI), on Landsat 8 and the MultiSpectral Instrument (MSI) on Sentinel-2. HLS products are generated using a set of algorithms that ensure seamless integration of data from both sensors, including atmospheric correction, cloud and cloud-shadow masking, spatial co-registration, common gridding, bidirectional reflectance distribution function normalisation

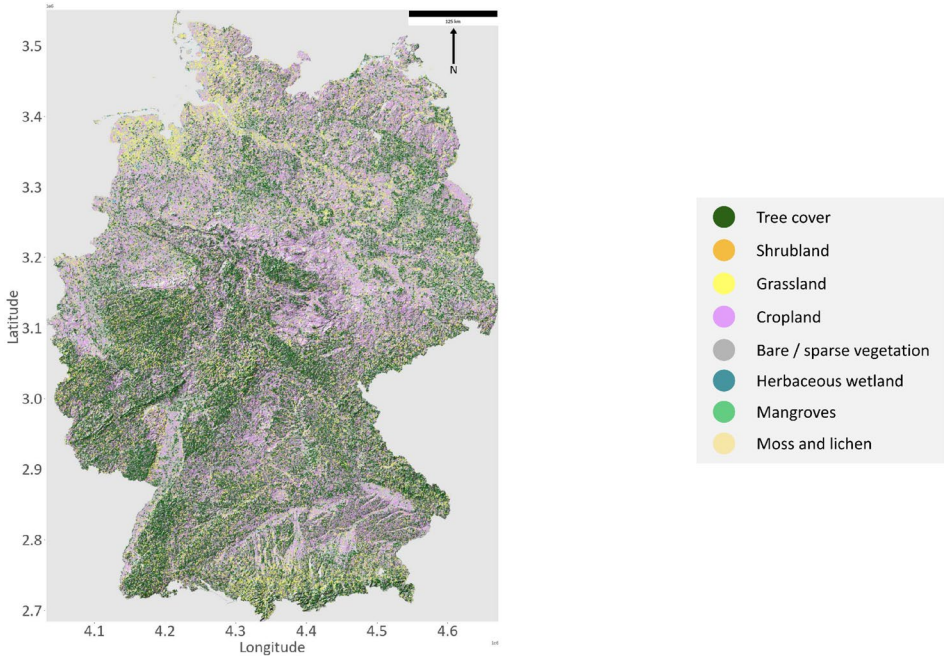

FIGURE 3 | Land cover classes in selected locations with corresponding legends.

TABLE 1 | Statistical summaries of SOC samples utilised in this study.

	Mean	s.d.	Min.	Q1	Median	Q3	Max.
LUCAS samples[a]	36.86	55.41	2.20	13.20	21.00	37.50	559.70

[a]The SOC content is measured in (g/kg).

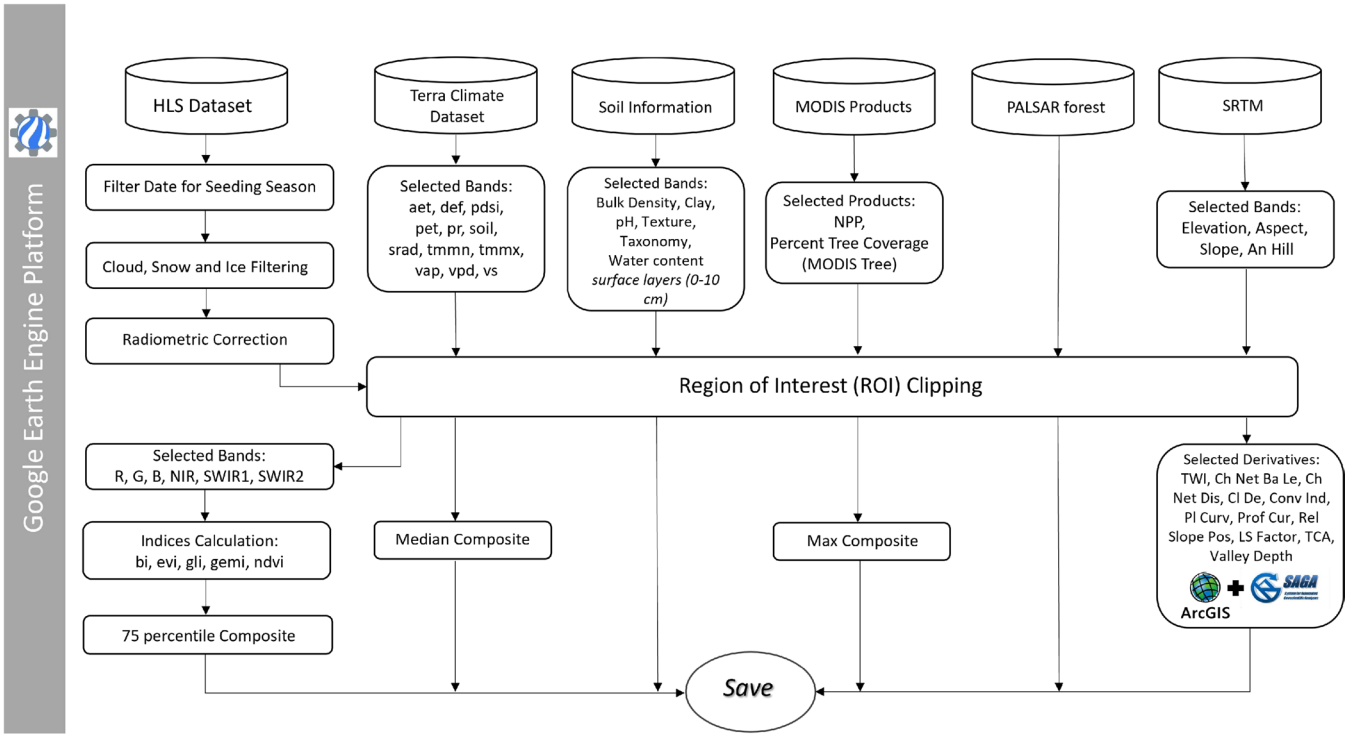

FIGURE 4 | The process of preparing input features in GEE for various applied products.

and spectral bandpass adjustment (Claverie et al. 2018). The HLS dataset provides a higher temporal resolution compared to individual Landsat-8 and Sentinel-2 datasets, which is particularly beneficial for our analysis, especially for removing cloud-affected images.

To minimise the impact of vegetation on our analysis, we focused on images captured during the seeding season as recommended by Dvorakova et al. (2023), specifically between day of year 70 and 120. This period is less likely to be affected by dense vegetation cover. We also applied a filtering process to exclude images with significant cloud cover, snow, ice and shadow, ensuring these accounted for less than 20% of the total image area. Given the high likelihood of cloud cover during this period in most European countries, we included a 5-year interval starting from 2014 to increase the probability of obtaining usable images. The final selected HLS bands are red, green, blue, near-infrared and two bands of shortwave infrared.

2.2.3 | Vegetation

Natural vegetation serves as one of the major sinks for terrestrial organic carbon (Razzaghi et al. 2022), making vegetation data crucial for predictive modelling. To assess information about natural vegetation, we utilise various features. First, we consider net primary productivity (NPP), which measures the rate at which plants absorb atmospheric carbon through the balance of photosynthesis and plant respiration (Yuan et al. 2021). Additionally, we gather data on tree cover percentage using a product from MODIS. For both features, we used the maximum composite for Germany. The other dataset we used is PALSAR, which shows forest and non-forest areas.

We also carefully examined different remote sensing indices and identified five key ones for our study. The Normalised Difference Vegetation Index (NDVI) is widely used for predicting soil properties. The Enhanced Vegetation Index (EVI) is similar to NDVI but corrects for atmospheric conditions and canopy background noise, making it more sensitive in areas with dense vegetation (Grunwald 2009). Additionally, we used the Global Environment Monitoring Index (GEMI), which is designed to minimise the effects of atmospheric disturbances while retaining information about vegetation cover (Pinty and Verstraete 1992). The Green Leaf Index (GLI) ranges from −1 to +1, with negative values representing soil and non-living features, and positive values representing green leaves and stems (Louhaichi et al. 2001). Finally, we used the Bare Soil Index (BI), a numerical indicator that captures soil variations (Rikimaru et al. 2002). The formulas for these indices are provided in Table 3.

2.2.4 | Climate

To investigate the diverse environmental conditions that either promote or impede climate regulation, we utilised a comprehensive set of parameters provided by TerraClimate (Abatzoglou et al. 2018). These parameters were derived from gridded meteorological data using a climatically aided spatio-temporal interpolation technique applied to the WorldClim datasets (Hijmans et al. 2005), enabling the estimation of monthly time series. The environmental factors selected for this study were divided into two categories: (1) primary climate variables, including maximum temperature, minimum temperature, vapour pressure, precipitation accumulation and downward surface shortwave radiation, and (2) derived or secondary variables, including reference evapotranspiration,

TABLE 2 | The features utilised in this study along with their descriptions.

No	Feature	Description	Unit	Category
1	Bulk density	Bulk density	Percent	Soil
2	Clay	Clay content	Percent	Soil
3	pH	pH	Unitless	Soil
4	Soil moisture	Soil moisture	mm	Soil
5	Texture	Soil texture (Class)	Unitless	Soil
6	Taxonomy	Taxonomy (Class)	Unitless	Soil
7	Water content	Water content	Percent	Soil
8	HLS bi	Harmonised LandSat-Sentinel Bare Soil Index	Unitless	Soil
9	HLS red	Harmonised LandSat-Sentinel red band	Unitless	RS
10	HLS green	Harmonised LandSat-Sentinel green band	Unitless	RS
11	HLS blue	Harmonised LandSat-Sentinel blue band	Unitless	RS
11	HLS nir	Harmonised LandSat-Sentinel near infrared band	Unitless	RS
12	HLS swri1	Harmonised LandSat-Sentinel shortwave infrared1 band	Unitless	RS
13	HLS swri2	Harmonised LandSat-Sentinel shortwave infrared2 band	Unitless	RS
14	HLS evi	Harmonised LandSat-Sentinel Enhanced Vegetation Index	Unitless	Vegetation
15	HLS gli	Harmonised LandSat-Sentinel Green Leaf Index	Unitless	Vegetation
16	HLS gemi	Harmonised LandSat-Sentinel Global Environment Monitoring Index	Unitless	Vegetation
17	HLS ndvi	Harmonised LandSat-Sentinel Normalized Vegetation Index	Unitless	Vegetation
18	MODIS Tree	The fraction (%) of pixels that are covered by trees	Unitless	Vegetation
19	PALSAR Forest	The global forest/non-forest map	Unitless	Vegetation
20	NPP	Net primary product MODIS (NPP), Terra sensor	$kg*C/m^2$	Vegetation
21	aet	Actual evapotranspiration, derived using a one-dimensional soil water balance model	mm	Climate
22	pdsi	Palmer drought severity index	Unitless	Climate
23	def	Climate water deficit, derived using a one-dimensional soil water balance model	mm	Climate
24	pet	Reference evapotranspiration	mm	Climate
25	pr	Precipitation accumulation	mm	Climate
26	srad	Downward surface shortwave radiation	W/m^2	Climate
27	tmmn	Minimum temperature	°C	Climate
28	tmmx	Maximum temperature	°C	Climate
29	vap	Vapour pressure deficit	kPa	Climate
30	vpd	Vapour pressure	kPa	Climate
31	vs	Wind speed at 10 m	m/s	Climate
32	An Hill	Analytical hillshading	Unitless	Topography
33	Elevation	Elevation	metres	Topography
34	Slope	Slope	Percent	Topography
35	Aspect	Aspect	Percent	Topography

(Continues)

TABLE 2 | (Continued)

No	Feature	Description	Unit	Category
36	TWI	Topographic wetness index	Unitless	Topography
37	Ch Net Ba Le	Channel network base level	metres	Topography
38	Ch Net Dis	Channel network distance	metres	Topography
39	Cl De	Closed depressions	Unitless	Topography
40	Conv Ind	Convergence index	Unitless	Topography
41	Pl Curv	Plan curvature	1/m	Topography
42	Prof Cur	Profile curvature	1/m	Topography
43	Rel Slope Pos	Relative slope position	Unitless	Topography
44	LS Factor	LS-factor	Unitless	Topography
45	TCA	Total catchment area	m^2	Topography
46	Valley Depth	Valley depth	metres	Topography

TABLE 3 | Remote sensing indices and their corresponding formulas.

Index	Formula
NDVI	$(\mathrm{NIR}-R)/(\mathrm{NIR}+R)$
EVI	$2.5\times((\mathrm{NIR}-R)/(\mathrm{NIR}+6\times R-7.5\times \mathrm{B}+1))$
GEMI	$A\times(1-0.25\times A)-(R-0.125)/(1-R)$, $A=\frac{(2\times((N^2)-(R^2))+1.5\times N+0.5\times R)}{(N+R+0.5)}$
GLI	$(2.0\times G-R-B)/(2.0\times G+R+B)$
BI	$((\mathrm{Swir1}+R)-(N+B))/((\mathrm{Swir1}+R)+(N+B))$

actual evapotranspiration, climate water deficit, soil moisture, Palmer drought severity index (PDSI) and vapour pressure deficit. These specific variables were chosen due to their well-recognised influence on SOC dynamics (Sakhaee et al. 2022; Fick and Hijmans 2017) and their role in the ecosystem's climate regulation function (Tamburini et al. 2020; Yang et al. 2020). Meteorological data from aforementioned 5 years were downloaded for this study and the median composite of the data were used for this study.

2.2.5 | Topography

We utilised digital elevation data from the Shuttle Radar Topography Mission (SRTM), specifically the SRTM-V3 (SRTM Plus) product provided by NASA JPL, with a resolution of 1 arc-second (approximately 30 m) (Farr et al. 2007). To incorporate topographic and geomorphological relief features, additional variables derived from this dataset were included, such as slope, aspect, topographic wetness index, profile curvature and valley depth. A comprehensive list of these variables is provided in Table 2. These topographic factors influence soil distribution across the landscape, impacting SOC dynamics through mechanisms such as overland flow and erosion (Carter and Ciolkosz 1991; Scholten et al. 2017).

3 | Methodology

3.1 | Explanation Model

The main idea is to create a separate explanation model, $\hat{f}_{\mathrm{expl}}$, to interpret the prediction model, $\hat{f}_{\mathrm{pred}}$, as shown in Figure 5. This flexible approach does not require retraining or modifying the prediction model to explain its outputs and has been thoroughly evaluated using well-known datasets, such as MNIST, ImageNet, Boston Housing and CIFAR10 (Schwab and Karlen 2019). We use an objective function to measure the importance of individual input features to the prediction model's accuracy, thereby training the explanation model. This method transforms the task of estimating feature relevance for a prediction model into a supervised learning problem, solvable with existing supervised ML models.

The core of the proposed explanation model is its objective function, which enables the optimization of the explanation model to elucidate another predictive model. The objective on which we base our work was first proposed to provide accurate predictions and estimates of feature importance in a single NN model (Schwab et al. 2019). This version does not depend on any specific model structure, allowing it to train explanation models to interpret any ML model. It is important to note that the objective from (Schwab et al. 2019) is grounded in Granger's definition of causality (Granger 1969). According to this principle, a relationship $X \to Y$ exists if we can predict Y more accurately with all available information than without X. In other words, the absence of x_i diminishes our ability to accurately predict $\hat{y}$. Given input features X, we define $\epsilon_{X\setminus\{i\}}$ as the predictive model's error when no information from the i th input feature is included and ϵ_X as the predictive model's error when all available input features are considered. We use the loss function of the predictive models $\mathrm{Loss}_{\mathrm{pred}}$ to compare the predictions to the ground reference values y and thus the aforementioned errors.

$$\Delta\epsilon_{X,i}=\epsilon_{X\setminus\{i\}}-\epsilon_X=\mathrm{Loss}_{\mathrm{pred}}\left(y,\hat{y}_{X\setminus\{i\}}\right)-\mathrm{Loss}_{\mathrm{pred}}\left(y,\hat{y}_X\right) \quad (1)$$

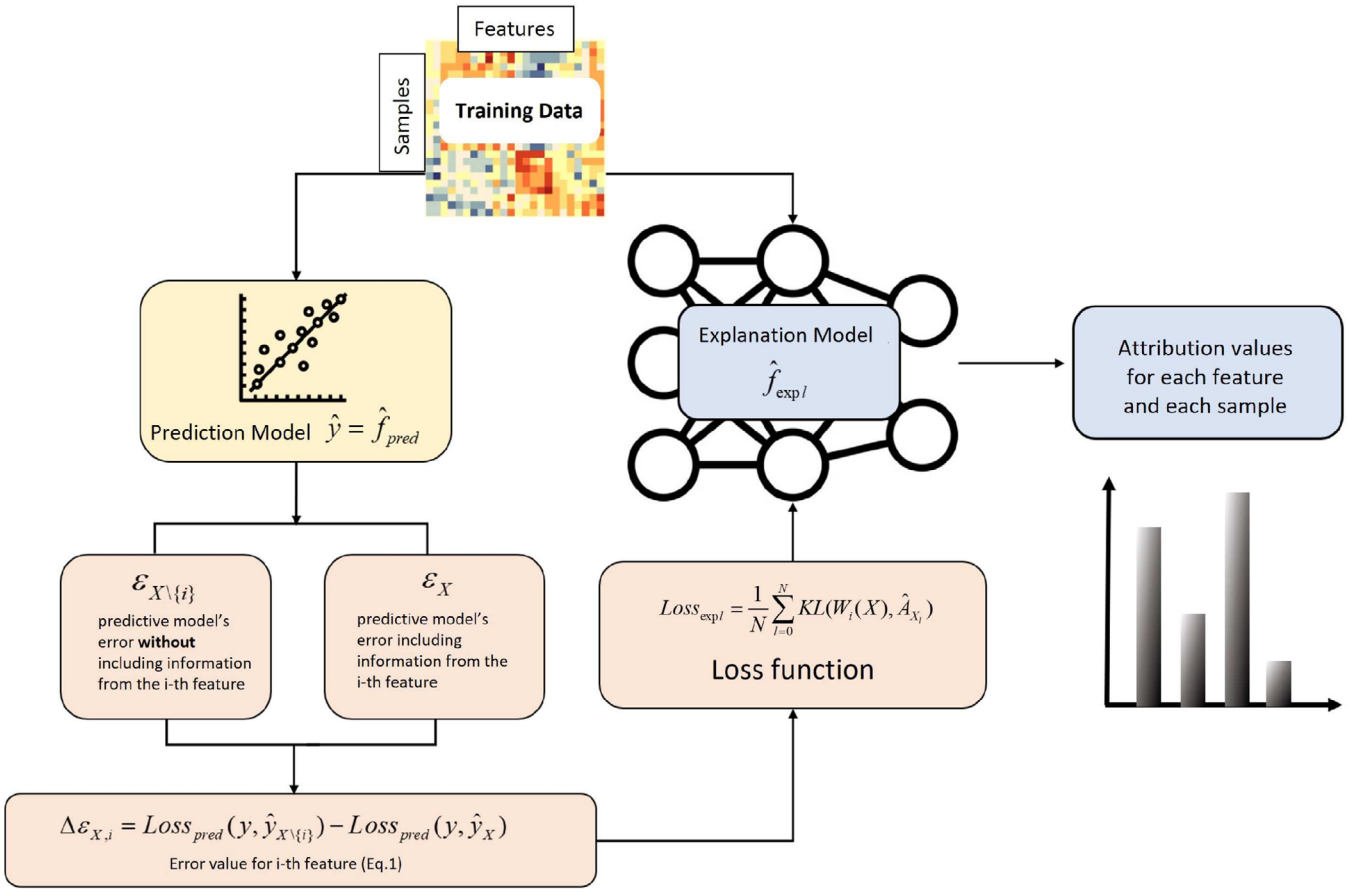

FIGURE 5 | Conceptual diagram of the proposed explanation model.

To quantify the contribution of the *i*th input feature to the model's output, denoted $\Delta\varepsilon_{X,i}$, we evaluate the reduction in prediction error when this feature is included, using the loss function $\text{Loss}_{\text{pred}}$, which takes the initial input features as inputs. This measure indicates the reduction in error. We implement a masking technique to elucidate the contributions of different features. This technique involves systematically replacing portions of the input data with zeros and observing the resulting changes in model outputs. The process includes dividing the input into manageable batches and applying zero imputation on each batch. For our explanation model, which is a two-layer multilayer perceptron (MLP), both the original and masked inputs are processed to obtain predictions, which are then aggregated. This process relies on four key components: the original inputs, their corresponding predictions, the masked inputs and the predictions for the masked data.

Finally, we define the objective function using normalised relative errors, $\Delta\varepsilon_{X,i}$, referred to as $W_i(X)$, and the importance scores, denoted $\widehat{A}$.

$$\text{Loss}_{\text{expl}} = \frac{1}{N}\sum_{j=1}^{N} KL\left(W_i(X), \widehat{A}_{X_j}\right) \tag{2}$$

In Equation (2), KL denotes the Kullback–Leibler (KL) divergence (Kullback 1997), which is expressed as follows:

$$\text{Loss}_{\text{expl}} = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{M} W_i(X)\log\left(\frac{W_i(X)}{\widehat{A}_{X_j}}\right) \tag{3}$$

where N represents the number of samples, M represents the number of features, $W_i(X)$ denotes the relative error for feature i, $\widehat{A}_{X_j}$ is the importance score for feature i of sample j.

To calculate $\widehat{A}$, which quantifies the importance of each feature, we employ an explanation model based on an MLP equipped with an embedded scoring mechanism. Initially, the input data are processed through the model, where each feature is assigned an initial score reflecting its contribution to the prediction. These scores are refined iteratively during training, allowing the model to identify the most influential features. To further assess feature importance, the MLP evaluates each feature by masking it individually to simulate its absence. This involves running the model twice for each feature: once with all features present and once with the target feature masked. The change in model error, calculated using KL divergence (Equation 3), indicates the feature's influence on the prediction—a greater shift in error reflects a higher feature importance. Finally, these refined scores are normalised with the softmax function, ensuring they sum to 1, creating a probability distribution that represents each feature's overall contribution, denoted $\widehat{A}$.

Minimising the function in Equation (3) aims to reduce the discrepancy between the distribution of feature importance in the training data, represented by W, and the learned importance scores $\widehat{A}$ for each sample X. This alignment ensures that the model's scoring mechanism accurately reflects the quantified contributions of individual features.

3.2 | Evaluation Metrics

When evaluating the performance of predictive models, several commonly used evaluation metrics include mean absolute error (MAE), the coefficient of determination (R^2), root mean square error (RMSE), ratio of performance to interquartile distance (RPIQ) and concordance correlation coefficient (CCC). MAE is calculated by taking the average of the absolute differences between the predicted values and the true values.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\, y_i - \hat{y}_i \,| \tag{4}$$

R^2 (also known as the coefficient of determination) is a statistical measure that represents the proportion of the variance in the dependent variable that can be explained by the independent variables in a regression model. It is commonly used to assess the goodness of fit of a regression model. This value is calculated using the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{5}$$

where n is the number of observations, y_i is the observed value of the dependent variable for observation i, $\hat{y}_i$ is the predicted value of the dependent variable for observation i based on the regression model and $\bar{y}$ is the mean of the observed values of the dependent variable.

RMSE is calculated by taking the square root of the average of the squared differences between the predicted values and the true values.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{6}$$

RPIQ represents the spread of the population and is calculated using the following equation (Bellon-Maurel et al. 2010):

$$\mathrm{RPIQ} = \frac{Q_3 - Q_1}{\mathrm{RMSE}} \tag{7}$$

The values Q_1 and Q_3 represent the 25th and 75th percentiles of the true samples, respectively, defining the interquartile distance.

CCC is a measure of the agreement between the predicted values and the true values. It considers both the mean difference and the variance difference between the predicted and true values.

$$\mathrm{CCC} = \frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + \left(\mu_y - \mu_{\hat{y}}\right)^2} \tag{8}$$

Here, represents the correlation coefficient between the predicted and observed values, σ_y and $\sigma_{\hat{y}}$ are the standard deviations of the observed and predicted values, respectively, and μ_y $\mu_{\hat{y}}$ are the means of the observed and predicted values, respectively. These metrics provide quantitative measures to assess the accuracy, correction, agreement and calibration of predictive models compared to the observed values.

4 | Results

4.1 | Prediction Models

We have selected five commonly used ML models in the field of soil science as prediction model. The first model is RF, a popular ML algorithm for regression tasks, especially in soil science (e.g., Szatmári et al. 2021). RF is a versatile ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees for regression tasks. Its ability to handle large datasets with higher dimensionality and the capability to estimate the importance of different features make it a favoured choice in various fields. The second model, deep forest (DeepForest), is an advanced version of RF, which is a decision tree ensemble method that requires fewer hyperparameters than deep learning models (Zhou and Feng 2019). DeepForest enhances the traditional RF by incorporating a cascade structure that enables deep representation learning. Each layer in the cascade comprises multiple random forests and completely random trees, and the output from one layer serves as the input features for the next. This architecture allows DeepForest to achieve high performance with fewer hyperparameter tuning requirements compared to deep learning models, while also maintaining the interpretability of ensemble methods. Additionally, DeepForest's ability to adaptively determine the depth of the cascade based on validation performance ensures a balance between model complexity and generalisation capability. This makes it a powerful and flexible approach for handling a wide range of predictive modelling tasks, especially when dealing with small- to medium-sized datasets. The subsequent model is gradient boosting (GB). This method constructs an additive model by employing a predetermined number of decision trees as weak learners or weak predictive models (Natekin and Knoll 2013). GB falls under the category of boosting techniques utilised in ML. It operates under the assumption that incorporating the most promising next model, when combined with previous models, will decrease the overall prediction error. Thus, defining the desired outcomes for this subsequent model is crucial for error minimization. The following model to be explored is the NN, applicable for regression problems. Here, a basic two-layer NN employing a mean square loss function is utilised. Lastly, ridge regression (Ridge) is employed. Ridge regression, also referred to as L_2 regularisation, is one of various regularisation methods for linear regression models. Hyperparameters for all models were determined using randomised grid search. Seventy-five percent of the data were used for training via stratified sample selection, where the SOC values were divided into distinct categories, and samples were selected to ensure that each category was proportionally represented in the training set.

To thoroughly evaluate the performance of our selected model, we implemented a simple baseline approach. Given the severe skewness of our dataset, we initially used the median of the available ground truth (GT) for predictions. We then calculated all evaluation metrics for this naive model. Although this baseline model is not expected to perform well, as indicated by an R^2

value of zero, it provides a useful benchmark for assessing the effectiveness of our selected models compared to the most basic prediction scenario.

The evaluation metrics (Table 4) show clear differences regarding the performance of the five ML methods presented. DeepForest demonstrates superior accuracy across RMSE, MAE and R^2, which is expected due to its integration of random forest and deep learning capabilities in a single model. The NN model achieves a comparable RMSE to RF and a higher R^2, although with a lower RPIQ. Notably, the CCC of the NN model surpasses those of other models. While the RF model produces acceptable results, it is less accurate than its more advanced counterpart, DeepForest. GB exhibits the lowest RMSE and R^2. Lastly, the ridge model achieves the highest RPIQ, despite sharing the same R^2 as RF and displaying the highest MAE among the models considered.

To enhance comprehension of the different model performance, we sampled over 200,000 random locations within Germany, which were subsequently examined thoroughly. We excluded areas identified as permanent water bodies or urban regions. We have used the ESA land cover product (Zanaga et al. 2021) for this purpose. Subsequently, we applied the trained prediction models to predict SOC at these random locations and visualised the results in Figure 6. The distribution of predicted SOC is largely consistent across all ML models, yet discernible differences are evident.

The spatial patterns of SOC contents are captured with variations in distribution across the five ML models (Figure 6). The low mountain ranges are clearly emphasised, showing the highest SOC contents. Differences are notable in the coastal areas along the North Sea and the Baltic Sea, where DeepForest, RF, GB and Ridge exhibit clear spatial variability, while NN shows only minor differences. The range of predicted SOC contents is smallest for DeepForest, RF and Ridge, with values between 0 and ~100 g of SOC per kilogram of soil. NN and GB show about twice to almost three times higher contents. Among all models, NN and GB yielded more realistic prediction ranges, clearly exceeding 0 to ~100 g/kg but still less than the actual range of 0 to ~500 g/kg, making them better than the other models. DeepForest performs slightly better than RF, and the Ridge model exhibits a similar range to the RF model. There are some samples with very high values in the northwestern part of Germany that NN fails to recognise correctly. Interestingly, the strong deviations between measured and modelled data affect all landscape areas and the spatial patterns remain the same. The differences between the landscape areas are most pronounced for Ridge, with SOC contents close to zero in large areas in the North German Plain to the north and west of the low mountain ranges as well as in the Upper Rhine Graben and the Cologne Bay. However, all models successfully captured hotspots of SOC values in central and southern Germany. DeepForest outperformed RF, allowing it to accurately recognise both very low and very high SOC values within a single model. The Ridge model had a significant shortcoming: it produced negative values that had to be converted to zero, failing to predict low SOC values accurately. These insights were not apparent from evaluation metrics alone. Understanding model performance is crucial as it directly impacts the explanation model.

TABLE 4 | Performance results for selected ML methods as prediction model.

Model	RMSE ↓	R^2(%) ↑	RPIQ ↑	MAE ↓	CCC ↑
DeepForest	**46.71**	**23**	0.66	**20.84**	0.39
NN	48.37	18	0.47	21.47	**0.40**
RF	48.86	16	0.57	22.21	0.29
GB	50.48	11	0.60	22.76	0.38
Ridge	49.03	16	**0.78**	25.11	0.30
Base	55.85	0	0.42	27.98	—

4.2 | Explanation for Predictions

After making predictions, we used the trained models to generate explanations based on Equation (2). Specifically, we trained a two-layer MLP to explain each model's predictions across different scenarios. This explanation model provided an importance score (percentage) for each sample, indicating feature importance. The scoring mechanism ensures that the sum of all feature scores equals 1, allowing us to rank features by their contribution to the model's performance. After calculating these scores, we identified the most and least important features by ranking them. We then examined the top-contributing features at every 200,000 sample locations to determine the key predictors for each model. Finally, the top 80% of contributing features and their cumulative importance scores in the selected locations were visualised, as shown in Figure 7.

The analysis in Figure 7 demonstrates that each model utilises a varying number of features. Since importance scores are represented in percentages, the different feature counts across the plots indicate the degree to which each model depends on specific features to generate predictions. Interesting observations emerged from this analysis. DeepForest and NN were the methods that utilised the highest number of input features for prediction, with 25 features contributing in DeepForest and 23 in NN. This indicates that NNs inherently capture various information for prediction. Despite this commonality, they differed in terms of the most important features. In the DeepForest model, topography and remote sensing information were predominantly important for prediction, whereas in the NN model, soil information, specifically pH and bulk density, was the most crucial. For the GB model, which is another decision tree-based model, topography information was the most significant category of features affecting the prediction, with remote sensing surface information also being important. A notable result was that the RF model relied mostly on elevation information for prediction, with 80% of the prediction based on only four features, and climate and soil information were not significantly utilised in this specific model. In the ridge regression model, more than 45% of the information needed for prediction was provided

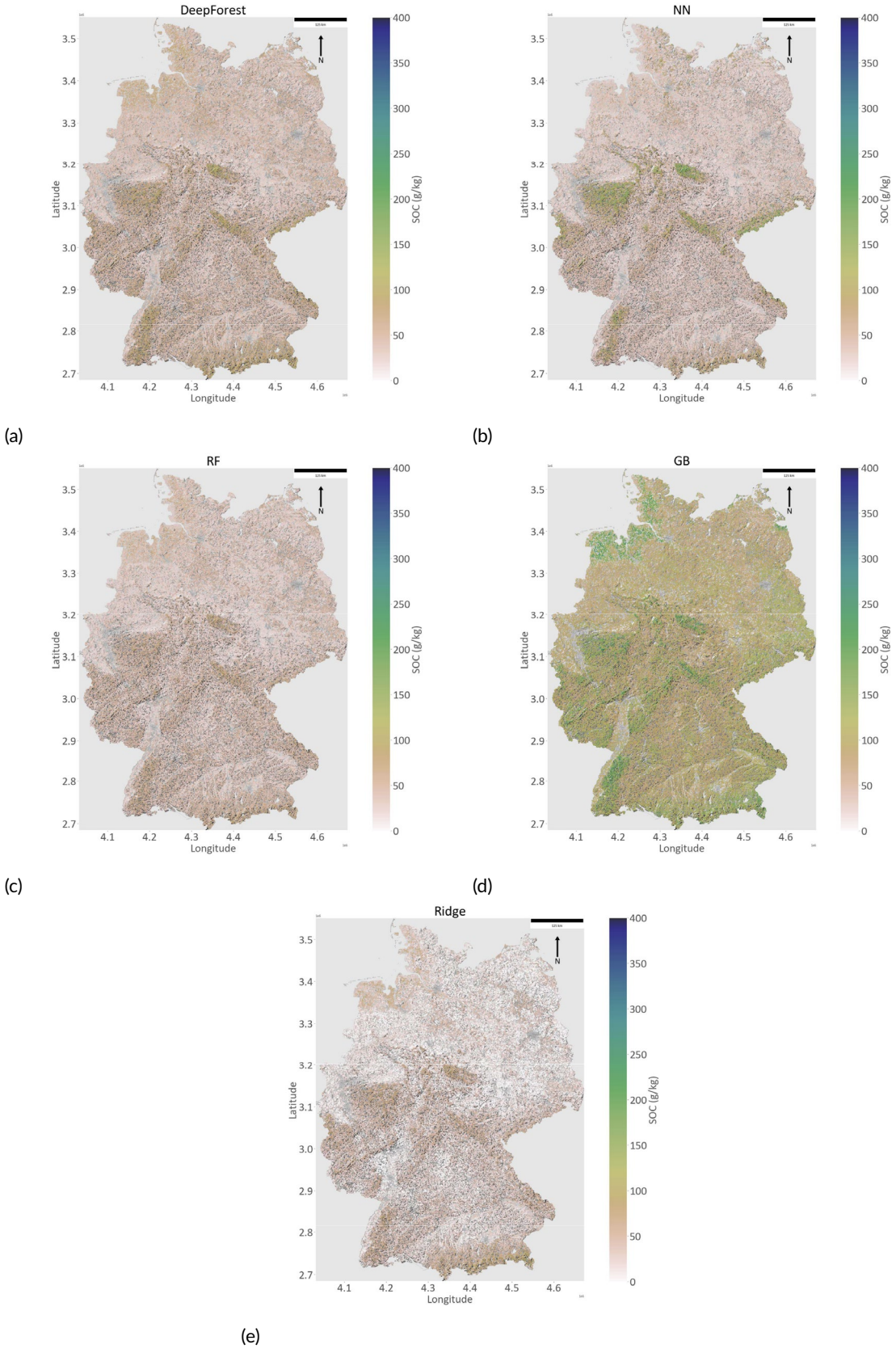

FIGURE 6 | The distribution of predicted SOC values in (g/kg) for selected random samples from (a) DeepForest, (b) NN, (c) RF, (d) GB and (e) Ridge model.

solely by soil information, specifically pH and water content. Overall, the most contributing features across all models were topography, soil and remote sensing (HLS) images. It is also important to note that soil information maps were primarily produced using surface information neglecting SOC stored in the subsoil, which can amount to about 50% of the organic C

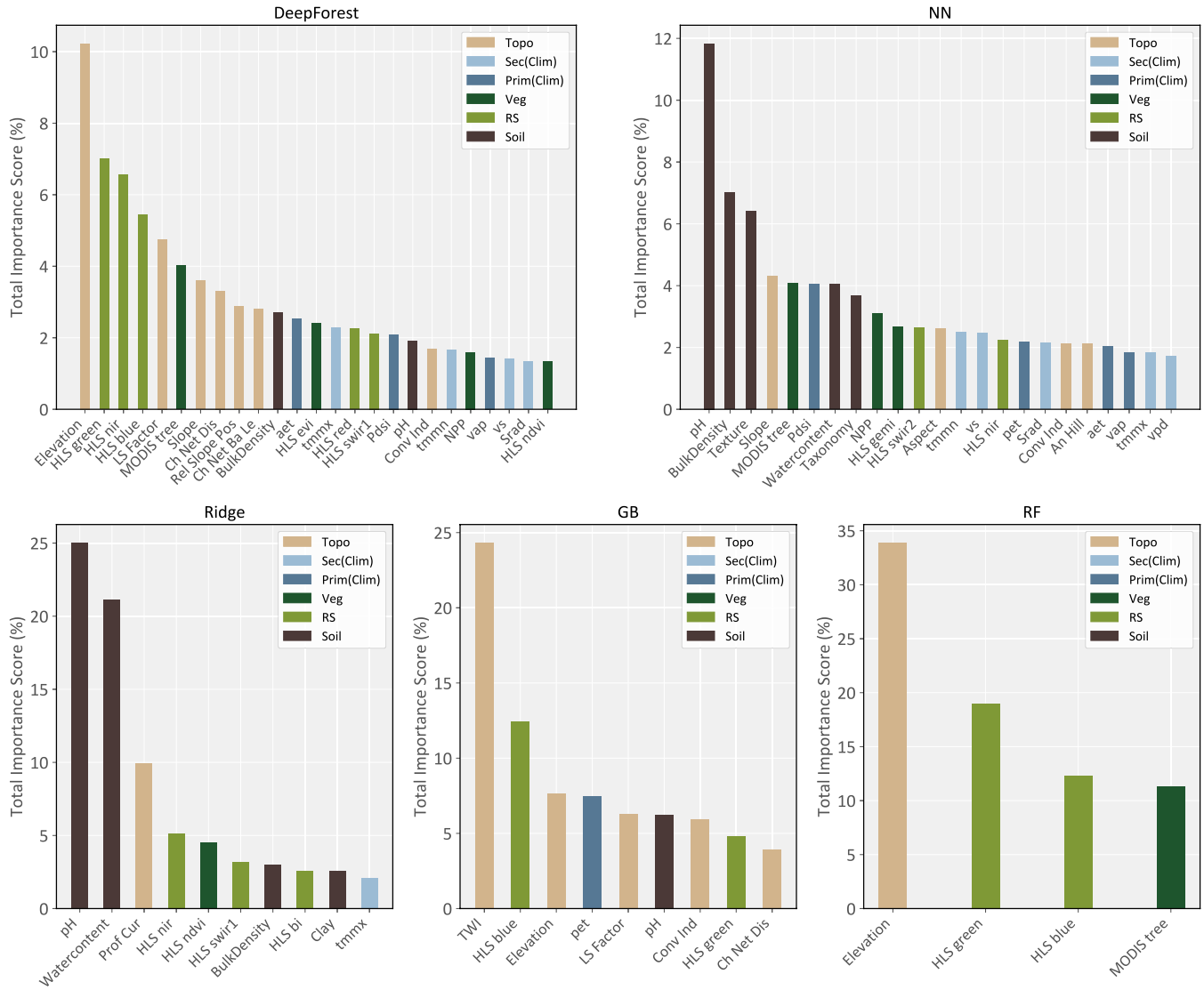

FIGURE 7 | The top 80% contributing features utilised for prediction across various models, derived by our proposed explanation model.

stored in soils worldwide (Batjes 1996). Finally, the percentage of importance attributed to each sample is the ultimate outcome of our explanation model. These outputs allow us to identify the highest importance score values and map the most important features at each location. This information is plotted in Figure 8. To ensure clarity, we marked only the categories in the plots.

4.3 | Comparison to Other Methods

In addition to the explanation model's calculation, model-specific feature importance can also be derived for RF and GB algorithms to identify key predictors. Ridge regression, which provides coefficients for each feature, also enables the assessment of feature significance. To compare the results with the explanation model, we plotted the model-specific feature importance for an 80% contribution threshold as well (Figure 9). For the RF model, the most important feature identified was elevation, consistent with our previous findings (Figure 7). The HLS green band and MODIS tree were ranked second and fourth, respectively, further confirming our results of the explanation model. Together with soil, these four features accounted for 80% of the predictions. However, the RF-specific feature importance highlights less importance score for these features and indicates that more features contribute to the predictions. It also does not specify their spatial locations or the exact percentage of their contributions. For the ridge regression model, the four most important features identified align with those found using the explanation model. However, the percentage contribution of these features calculated by our model is higher than those derived from the provided coefficient values.

In contrast, the feature importance for the GB model presents a different scenario. Although elevation remains a key predictor, the ranking of important features moves away from RS to vegetation and soil. GB constructs an ensemble of weak learners (i.e., decision trees) sequentially, with each learner addressing the errors of its predecessors. This iterative method aims to minimise the model's loss function, enhancing overall performance. Feature importance in GB is determined by each feature's contribution to reducing the loss function during training. On the other hand, our proposed approach

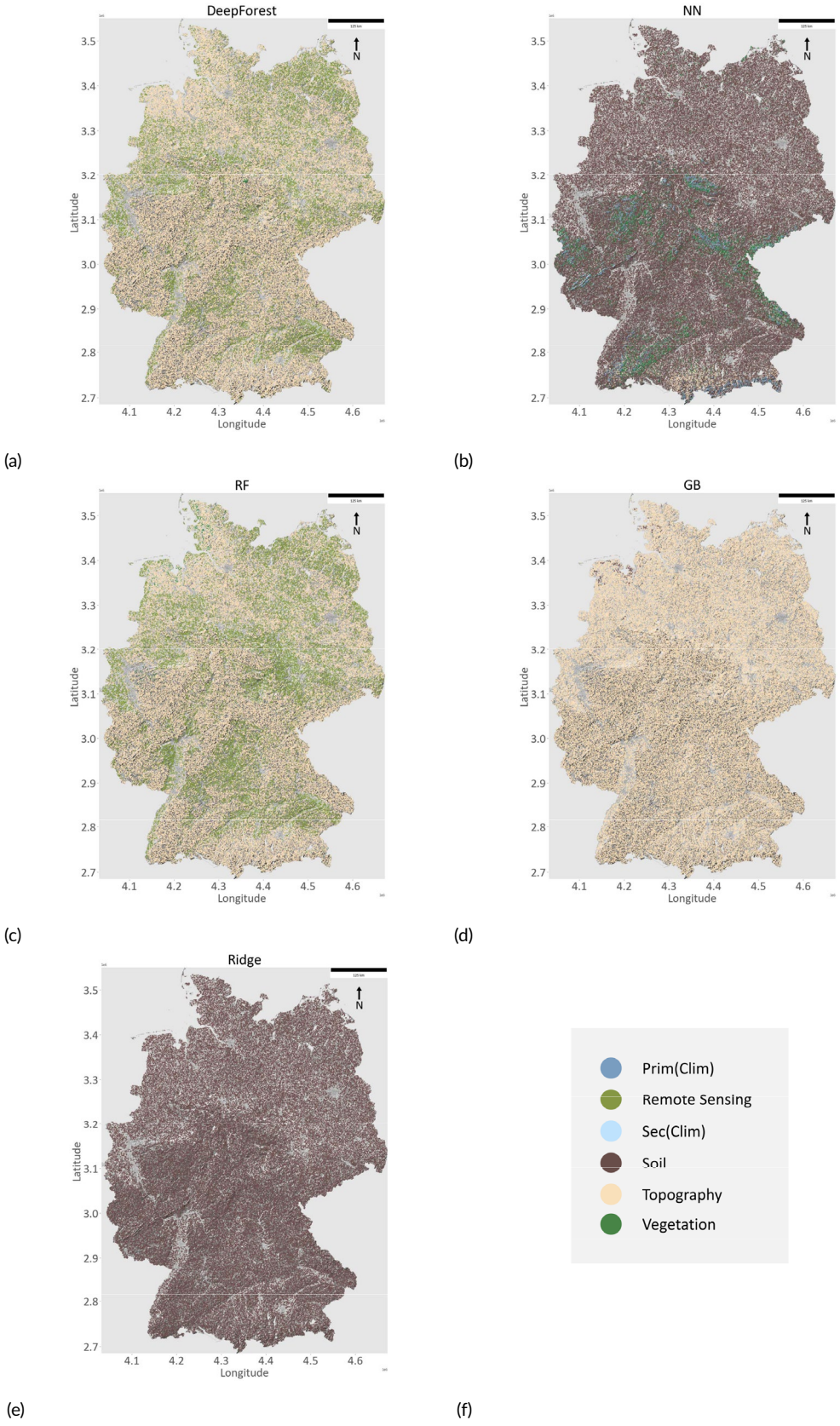

FIGURE 8 | The primary feature category utilised for prediction by each model at every location: (a) DeepForest, (b) NN, (c) RF, (d) GB, (e) Ridge model and (f) legend.

tries to find the importance scores of each feature based on the loss function of the trained prediction model after the training phase. However, it is not limited to only training samples; it can also be applied to any arbitrary samples. The explanation from our model shows that when used on larger samples, the feature importance scores can differ from model-specific

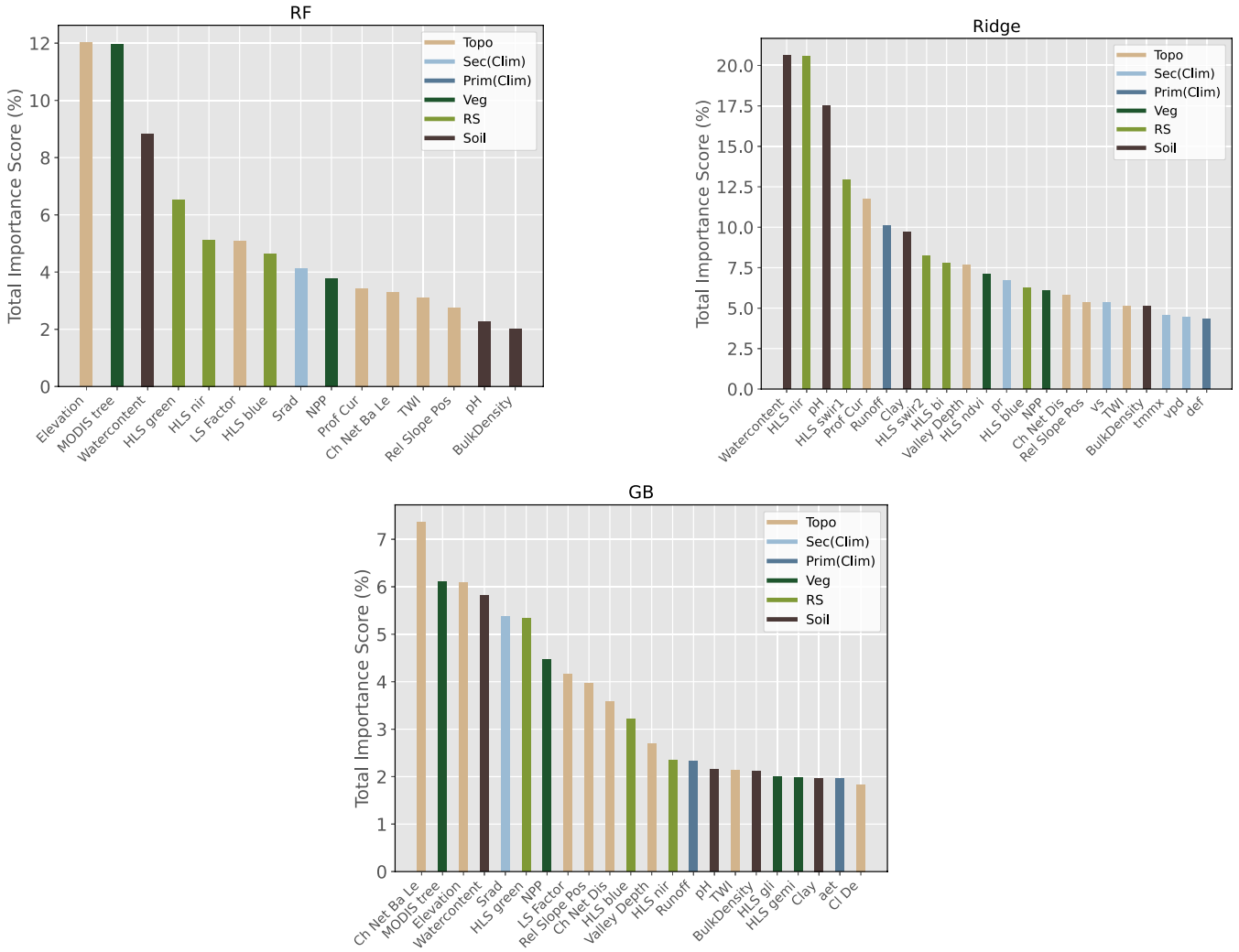

FIGURE 9 | The top 80% contributing features derived from model specific feature importance.

feature importance. One reason for this could be the lack of sufficient GT samples as training data. This affects the prediction model and, consequently, the explanation model because predictions for out-of-distribution samples, which are not reflected in the GT set, differ. As a result, the explanations we found for the training samples are not consistent with those for larger or different sample sets.

5 | Discussion

An immediate observation from the results in Figure 8 is that there is a significant difference between decision-tree-based models, such as DeepForest, RF and GB, and other models such as NN and ridge regression. The primary variable category used for prediction in the decision-tree-based models is topography. Topography plays a crucial role in SOC modelling, influencing the spatial distribution and variability of SOC across landscapes. Studies have shown that incorporating topographic variables into SOC models significantly enhances their accuracy and predictive power. For instance, Wiesmeier et al. (2013) demonstrated that topographic factors such as slope, aspect and elevation are key determinants of SOC distribution in agricultural soils in Bavaria, Germany. Similarly, Grunwald (2009) emphasised the importance of digital elevation models in capturing topographic variation, which is critical for accurate SOC predictions. Another study by Liu et al. (2003) highlighted that topography-driven water flow and erosion processes are essential for understanding SOC dynamics in heterogeneous landscapes. Similarly, a study on the spatial prediction of organic carbon in German agricultural topsoil highlighted the necessity of including topographic covariates to address the high variability of SOC (Sakhaee et al. 2022). These findings collectively underscore the necessity of including topographic information in SOC modelling to improve the reliability and robustness of predictions, which are vital for effective environmental management and carbon sequestration strategies. Another finding is the significant importance of surface information, specifically HLS composite images, for the performance of DeepForest and RF models. As shown in Figure 4, which examines land cover classes in selected locations, remote sensing information is particularly important in cropland areas (Broeg et al. 2024). This indicates that when utilising random forest models, which are widely used in the DSM field, meticulous preparation of RS data is crucial for accurate SOC prediction.

NN and Ridge model explanations indicate that these models predominantly rely on soil information such as pH, bulk

density, texture class and water content (see Figure 7). Each of these variables plays a critical role in predicting SOC and enhancing the accuracy of SOC models. Soil pH significantly influences microbial activity and organic matter decomposition rates, directly affecting SOC storage. High pH can promote the stability of organic matter, while low pH conditions may accelerate decomposition and carbon loss. A study by Kemmitt et al. (2006) highlighted the critical role of soil pH in controlling microbial processes and organic matter dynamics. Soil acidity, given as pH values, is an important feature for predicting SOC contents in topsoils of Germany, which was to be expected for large-scale observations (Luo et al. 2017). Topsoils rich in organic matter produce more organic acids during the decomposition of organic compounds (Hong et al. 2019). Other well-known processes that produce protons in soils and hinder the decomposition of SOC include the formation and dissociation of H^+ ions from carbonic acids, silicate weathering, nitrification, microbial respiration and the release of plant root exudates.

Bulk density is a measure of soil compaction, which affects porosity and root penetration. As noted in Franzluebbers (2002), bulk density is inversely related to SOC levels, making it a crucial variable in SOC modelling. Soil texture class, which includes the proportions of sand, silt and clay, determines soil structure and influences water retention and aeration. Clay soils, for example, can protect organic matter from decomposition due to their fine particles and strong aggregation, leading to higher SOC content. This relationship is thoroughly examined in the work of Hassink (1997), who demonstrated that finer-textured soils tend to have higher SOC due to better protection mechanisms. Water content in soil impacts microbial activity and organic matter decomposition. Adequate moisture levels promote microbial processes that contribute to SOC accumulation, while excessive or insufficient moisture can hinder these processes. The influence of soil moisture on SOC is extensively discussed in Xu et al. (2014), which emphasised that water content is pivotal for understanding soil respiration and carbon fluxes. These studies underscore the importance of incorporating soil pH, bulk density, texture class and water content into SOC models to improve their predictive accuracy and reliability.

Another interesting finding, from a mathematical perspective, emerged when examining the NN explanation. The model prioritised input features such as climate and vegetation in central parts of Germany and high elevation areas in the Alpine region. When observing the SOC map derived from this model in Figure 6b, we notice that these areas exhibit high variability and heterogeneity in SOC values, making it challenging for the model to predict SOC values accurately for samples in these regions. This difficulty may arise because the gradient of the loss function for the weights that are associated with soil information features is not sufficiently minimised. During the training process, the optimization algorithm adjusts the weights based on the gradients of the loss function. Features that contribute more to reducing the loss will have their corresponding weights updated more significantly (Goodfellow et al. 2016). This suggests that although in most areas, there are a limited number of highly influential features for an NN model, providing additional information to the NN could improve prediction accuracy in challenging areas where the data are heterogeneous. Ridge regression is the only model that predominantly relies on a single category, primarily pH and water content (see Figure 7). As a linear model, it fails to capture the nonlinearity within the data, leading it to depend solely on information that shows a high correlation with the dependent variable, SOC (Lukman et al. 2021).

Environmental factors significantly influence SOC modelling, but the specific variables affecting model accuracy differ across various ML models. Our findings demonstrate this variability, indicating that influencing factors are inconsistent across all models and depend on each model's inherent characteristics and design. Therefore, interpreting data-driven studies requires a meticulously structured approach, guided by expert knowledge and a comprehensive understanding of the model under investigation, as suggested by Runge et al. (2019). It is important to emphasise that our explanation model provides a comprehensive framework for understanding the underlying patterns and behaviours within the data, offering significant theoretical advancements. Our findings were efficiently obtained through the specific design of our proposed approach. In contrast, methods such as SHAP rely on approximations to manage importance calculations, which can introduce inaccuracies and are often computationally intensive, especially with complex models and large datasets. However, its practical application in real methodological decisions necessitates further integration with accuracy and uncertainty analyses. This additional step is essential to ensure that the model not only enhances interpretability but also maintains reliability and precision in practical scenarios. Additionally, the impact of feature dependence on interpretability, as highlighted by recent studies (Aas et al. 2021; Heskes et al. 2020), warrants attention. Future research will focus on integrating causal knowledge and dependency-aware techniques to further refine the model's explanations, ensuring they remain robust even when features exhibit interdependencies. Consequently, future research will focus on applying this model in conjunction with rigorous accuracy and uncertainty assessments to validate its effectiveness in real-world applications.

6 | Conclusions

We proposed and implemented an explanation model for ML models, using SOC content as GT and a wide range of environmental features as predictors. This explanation model frames the task of explaining ML model decisions as a learning problem, training the explanation model to assess the extent to which specific inputs influence the outputs of another ML model. The explanations provided insights into the features influencing SOC content variation and how the model's predictions were altered. Our explanation model is straightforward to employ, as it does not require retraining or modifying the original model. It demonstrates significant potential in interpreting complex ML models, particularly in soil science. The explanation model revealed not only the importance score of environmental features to SOC prediction but also the spatial pattern of feature contributions. The following findings can be inferred:

- Topography is the primary feature influencing the prediction ability of decision tree models, regardless of their complexity. Therefore, providing more accurate topographic information with higher spatial resolution could

significantly improve predictions in these family of ML models.

- Ridge regression models primarily rely on soil information due to their linear nature, which cannot capture non-linear relationships and depends heavily on variables that are highly correlated with the target variable.
- Providing a comprehensive set of environmental features can improve prediction accuracy in challenging areas with heterogeneous data for NNs.
- Environmental features affect SOC modelling differently across various ML models. Each model's unique characteristics determine which features are important.

Author Contributions

Nafiseh Kakhani: conceptualization, investigation, writing – original draft, methodology, validation, visualization, software, formal analysis, project administration, data curation. **Ruhollah Taghizadeh-Mehrjardi:** investigation, supervision. **Davoud Omarzadeh:** validation, visualization, writing – review and editing. **Masahiro Ryo:** writing – review and editing, validation, formal analysis. **Uta Heiden:** writing – review and editing. **Thomas Scholten:** resources, supervision, funding acquisition, writing – original draft, writing – review and editing, validation, formal analysis, investigation.

Acknowledgements

During the preparation of this work, the author(s) used ChatGPT-4 to assist with text editing. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication. This research was funded by the Deutsche Forschungsgemeinschaft (DFG) [3150] for the project 'MLTRANS-Transferability of Machine Learning Models in Digital Soil Mapping' and a grant to Thomas Scholten (SCHO 739/21–1) and 'Machine Learning for Science', which is part of Germany's Excellence Strategy—EXC number 2064/1—Project number 390727645. Davoud Omarzadeh is supported by a PhD grant from the Universitat Oberta de Catalunya (UOC). Open Access funding enabled and organized by Projekt DEAL.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

Aas, K., M. Jullum, and A. Løland. 2021. "Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values." *Artificial Intelligence* 298: 103502.

Abatzoglou, J. T., S. Z. Dobrowski, S. A. Parks, and K. C. Hegewisch. 2018. "TerraClimate, A High-Resolution Global Dataset of Monthly Climate and Climatic Water Balance From 1958–2015." *Scientific Data* 5: 1–12.

Ballabio, C., P. Panagos, and L. Monatanarella. 2016. "Mapping Topsoil Physical Properties at European Scale Using the LUCAS Database." *Geoderma* 261: 110–123.

Batjes, N. H. 1996. "Total Carbon and Nitrogen in the Soils of the World." *European Journal of Soil Science* 47: 151–163.

Bellon-Maurel, V., E. Fernandez-Ahumada, B. Palagos, J. M. Roger, and A. McBratney. 2010. "Critical Review of Chemometric Indicators Commonly Used for Assessing the Quality of the Prediction of Soil Attributes by NIR Spectroscopy." *TrAC Trends in Analytical Chemistry* 29, no. 9: 1073–1081.

Beugnon, R., W. Bu, H. Bruelheide, et al. 2023. "Abiotic and Biotic Drivers of Tree Trait Effects on Soil Microbial Biomass and Soil Carbon Concentration." *Ecological Monographs* 93: e1563.

Broeg, T., A., Don, A., Gocht, T., Scholten, R., Taghizadeh-Mehrjardi, and S., Erasmi. 2024. "Using Local Ensemble Models and Landsat Bare Soil Composites for Large-Scale Soil Organic Carbon Maps in Cropland." *Geoderma* 444: 116850.

Burkart, N., and M. F. Huber. 2021. "A Survey on the Explainability of Supervised Machine Learning." *Journal of Artificial Intelligence Research* 70: 245–317.

Carter, B. J., and E. J. Ciolkosz. 1991. "Slope Gradient and Aspect Effects on Soils Developed From Sandstone in Pennsylvania." *Geoderma* 49: 199–213.

Chen, J., S. Wu, A. Gupta, and R. Ying. 2024. "D4explainer: In-Distribution Explanations of Graph Neural Network via Discrete Denoising Diffusion." *Advances in Neural Information Processing Systems* 36: 1–12.

Chen, S., N. Kalanat, Y. Xie, et al. 2023. "Physics-Guided Machine Learning From Simulated Data With Different Physical Parameters." *Knowledge and Information Systems* 65, no. 8: 3223–3250.

Chuang, Y. N., G. Wang, F. Yang, et al. 2023. "Cortx: Contrastive Framework for Real-Time Explanation," arXiv preprint arXiv: 230302794.

Claverie, M., J. Ju, J. G. Masek, et al. 2018. "The Harmonized Landsat and Sentinel-2 Surface Reflectance Data Set." *Remote Sensing of Environment* 219: 145–161.

Covert, I., S. M. Lundberg, and S. I. Lee. 2021. "Explaining by Removing: A Unified Framework for Model Explanation." *Journal of Machine Learning Research* 22: 201–209.

Dvorakova, K., U. Heiden, K. Pepers, G. Staats, G. van Os, and B. van Wesemael. 2023. "Improving Soil Organic Carbon Predictions From a Sentinel–2 Soil Composite by Assessing Surface Conditions and Uncertainties." *Geoderma* 429: 116128.

Farr, T. G., P. A. Rosen, E. Caro, et al. 2007. "The Shuttle Radar Topography Mission." *Reviews of Geophysics* 45: 45.

Fick, S. E., and R. J. Hijmans. 2017. "WorldClim 2: New 1-Km Spatial Resolution Climate Surfaces for Global Land Areas." *International Journal of Climatology* 37: 4302–4315.

Franzluebbers, A. 2002. "Soil Organic Matter Stratification Ratio as an Indicator of Soil Quality." *Soil and Tillage Research* 66, no. 2: 95–106.

Gerke, H. H., H. Vogel, T. K. D. Weber, W. M. van der Meij, and T. Scholten. 2022. "3–4D Soil Model as Challenge for Future Soil Research: Quantitative Soil Modeling Based on the Solid Phase." *Journal of Plant Nutrition and Soil Science* 185: 720–744.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press.

Granger, C. W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica: Journal of the Econometric Society* 37, no. 3: 424–438. https://doi.org/10.2307/1912791.

Grunwald, S. 2009. "Multi-Criteria Characterization of Recent Digital Soil Mapping and Modeling Approaches." *Geoderma* 152, no. 3–4: 195–207.

Hassink, J. 1997. "The Capacity of Soils to Preserve Organic C and N by Their Association With Clay and Silt Particles." *Plant and Soil* 191: 77–87.

Hengl, T., J. M. de Jesus, G. B. M. Heuvelink, et al. 2017. "SoilGrids250m: Global Gridded Soil Information Based on Machine Learning." *PLoS One* 12: e0169748.

Heskes, T., E. Sijben, I. G. Bucur, and T. Claassen. 2020. "Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models." *Advances in Neural Information Processing Systems* 33: 4778–4789.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. "Very High Resolution Interpolated Climate Surfaces for Global Land Areas." *International Journal of Climatology: A Journal of the Royal Meteorological Society* 25, no. 15: 1965–1978.

Hong, S., P. Gan, and A. Chen. 2019. "Environmental Controls on Soil pH in Planted Forest and Its Response to Nitrogen Deposition." *Environmental Research* 172: 159–165.

Hostallero, D. E., L. Wei, L. Wang, J. Cairns, and A. Emad. 2023. "Preclinical-To-Clinical Anti-Cancer Drug Response Prediction and Biomarker Identification Using TINDL." *Genomics, Proteomics & Bioinformatics* 21, no. 3: 535–550.

Kemmitt, S. J., D. Wright, K. W. Goulding, and D. L. Jones. 2006. "pH Regulation of Carbon and Nitrogen Dynamics in Two Agricultural Soils." *Soil Biology and Biochemistry* 38, no. 5: 898–911.

Kullback, S. 1997. *Information Theory and Statistics*. Courier Corporation.

Lal, R. 2004. "Soil Carbon Sequestration Impacts on Global Climate Change and Food Security." *Science* 304: 1623–1627.

Liu, S., N. Bliss, E. Sundquist, and T. G. Huntington. 2003. "Modeling Carbon Dynamics in Vegetation and Soil Under the Impact of Soil Erosion and Deposition." *Global Biogeochemical Cycles* 17, no. 2: 2002GB002010. https://doi.org/10.1029/2002GB002010.

Louhaichi, M., M. M. Borman, and D. E. Johnson. 2001. "Spatially Located Platform and Aerial Photography for Documentation of Grazing Impacts on Wheat." *Geocarto International* 16, no. 1: 65–70.

Lukman, A. F., I. Dawoud, B. G. Kibria, Z. Y. Algamal, and B. Aladeitan. 2021. "A New Ridge-Type Estimator for the Gamma Regression Model." *Scientifica* 2021: 1–8.

Lundberg, S. M., and S. I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30: 4768–4777.

Luo, Z., W. Feng, Y. Luo, J. Baldock, and E. Wang. 2017. "Soil Organic Carbon Dynamics Jointly Controlled by Climate, Carbon Inputs, Soil Properties and Soil Carbon Fractions." *Global Change Biology* 23: 4430–4439.

Malik, A. A., J. Puissant, K. M. Buckeridge, et al. 2018. "Land Use Driven Change in Soil pH Affects Microbial Carbon Cycling Processes." *Nature Communications* 9: 1–10.

Mulder, V. L., M. Lacoste, M. P. Martin, A. Richer-de Forges, and D. Arrouays. 2015. "Understanding Large-Extent Controls of Soil Organic Carbon Storage in Relation to Soil Depth and Soil-Landscape Systems." *Global Biogeochemical Cycles* 29, no. 8: 1210–1229.

Natekin, A., and A. Knoll. 2013. "Gradient Boosting Machines, a Tutorial." *Frontiers in Neurorobotics* 7: 21.

Orgiazzi, A., C. Ballabio, P. Panagos, A. Jones, and O. Fernández-Ugalde. 2018. "LUCAS Soil, the Largest Expandable Soil Dataset for Europe: A Review." *European Journal of Soil Science* 69, no. 1: 140–153.

Pinty, B., and M. Verstraete. 1992. "GEMI: A Non-Linear Index to Monitor Global Vegetation From Satellites." *Vegetatio* 101: 15–20.

Powlson, D. S., A. P. Whitmore, and K. W. Goulding. 2011. "Soil Carbon Sequestration to Mitigate Climate Change: A Critical Re-Examination to Identify the True and the False." *European Journal of Soil Science* 62, no. 1: 42–55.

Razzaghi, S., K. R. Islam, and I. A. M. Ahmed. 2022. "Deforestation Impacts Soil Organic Carbon and Nitrogen Pools and Carbon Lability Under Mediterranean Climates." *Journal of Soils and Sediments* 22, no. 9: 2381–2391.

Rikimaru, A., P. Roy, S. Miyatake, et al. 2002. "Tropical Forest Cover Density Mapping." *Tropical Ecology* 43, no. 1: 39–47.

Roscher, R., B. Bohn, M. F. Duarte, and J. Garcke. 2020. "Explainable Machine Learning for Scientific Insights and Discoveries." *IEEE Access* 8: 42200–42216.

Runge, J., S. Bathiany, E. Bollt, et al. 2019. "Inferring Causation From Time Series in Earth System Sciences." *Nature Communications* 10: 1–13.

Sakhaee, A., A. Gebauer, M. Ließ, and A. Don. 2022. "Spatial Prediction of Organic Carbon in German Agricultural Topsoil Using Machine Learning Algorithms." *Soil* 8: 587–604.

Samek, W., T. Wiegand, and K. R. Müller. 2017. "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," arXiv preprint arXiv:170808296.

Scholten, T., P. Goebes, P. Kühn, et al. 2017. "On the Combined Effect of Soil Fertility and Topography on Tree Growth in Subtropical Forest Ecosystems—A Study From SE China." *Journal of Plant Ecology* 10: 111–127.

Schwab, P., and W. Karlen. 2019. "CXPlain: Causal Explanations for Model Interpretation Under Uncertainty." In *Advances in Neural Information Processing Systems (NeurIPS) 2019 Conference Proceedings*. 10220–10230.

Schwab, P., D. Miladinovic, and W. Karlen. 2019. "Granger-Causal Attentive Mixtures of Experts: Learning Important Features With Neural Networks." *Proceedings of the AAAI Conference on Artificial Intelligence* 33, no. 1: 4846–4853. https://doi.org/10.1609/aaai.v33i01.33014846.

Situ, X., I. Zukerman, C. Paris, S. Maruf, and G. Haffari. 2021. "Learning to Explain: Generating Stable Explanations Fast." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol. 1, 5340–5355. Long Papers.

Smith, P., D. Martino, Z. Cai, et al. 2008. "Greenhouse Gas Mitigation in Agriculture." *Philosophical Transactions of the Royal Society, B: Biological Sciences* 363, no. 1492: 789–813.

Sundararajan, M., and A. Najmi. 2020. "The Many Shapley Values for Model Explanation." In *Proceedings of the 37th International Conference on Machine Learning*, edited by H. Daumé III and A. Singh, (Vol. 119, pp. 9269–9278. PMLR).

Szatmári, G., L. Pásztor, and G. B. Heuvelink. 2021. "Estimating Soil Organic Carbon Stock Change at Multiple Scales Using Machine Learning and Multivariate Geostatistics." *Geoderma* 403: 115356.

Tamburini, G., R. Bommarco, T. C. Wanger, et al. 2020. "Agricultural Diversification Promotes Multiple Ecosystem Services Without Compromising Yield." *Science Advances* 6: eaba1715.

Toth, G., A. Jones, L. Montanarella, et al. 2013. *LUCAS Topoil Survey - Methodology, Data and Results. EUR 26102*. Publications Office of the European Union.

Tziolas, N., N. Tsakiridis, U. Heiden, and B. van Wesemael. 2024. "Soil Organic Carbon Mapping Utilizing Convolutional Neural Networks and Earth Observation Data, a Case Study in Bavaria State Germany." *Geoderma* 444: 116867.

Vos, C., A. Don, E. U. Hobley, R. Prietz, A. Heidkamp, and A. Freibauer. 2019. "Factors Controlling the Variation in Organic Carbon Stocks in Agricultural Soils of Germany." *European Journal of Soil Science* 70: 550–564.

Vowels, M. J. 2022. "Trying to Outrun Causality With Machine Learning: Limitations of Model Explainability Techniques for Identifying Predictive Variables." *Stat* 22: 1050.

Wadoux, A. M. C. 2023. "Interpretable Spectroscopic Modelling of Soil With Machine Learning." *European Journal of Soil Science* 74, no. 3: e13370.

Wadoux, A. M. C., B. Minasny, and A. B. McBratney. 2020. "Machine Learning for Digital Soil Mapping: Applications, Challenges and Suggested Solutions." *Earth-Science Reviews* 210: 103359.

Wadoux, A. M. C., and C. Molnar. 2022. "Beyond Prediction: Methods for Interpreting Complex Models of Soil Variation." *Geoderma* 422: 115953.

Wadoux, A. M. C., N. Saby, and M. P. Martin. 2022. *Shapley Values Reveal the Drivers of Soil Organic Carbon Stocks Prediction*, 1–25. EGUsphere.

Wang, X., J. Han, X. Wang, H. Yao, and L. Zhang. 2021. "Estimating Soil Organic Matter Content Using Sentinel-2 Imagery by Machine Learning in Shanghai." *IEEE Access* 9: 78215–78225.

Wiesmeier, M., R. Hübner, F. Barthold, et al. 2013. "Amount, Distribution and Driving Factors of Soil Organic Carbon and Nitrogen in Cropland and Grassland Soils of Southeast Germany (Bavaria)." *Agriculture, Ecosystems & Environment* 176: 39–52.

Xu, X., J. P. Schimel, P. E. Thornton, X. Song, F. Yuan, and S. Goswami. 2014. "Substrate and Environmental Controls on Microbial Assimilation of Soil Organic Carbon: A Framework for Earth System Models." *Ecology Letters* 17, no. 5: 547–555.

Xu, X., X. Wang, P. Zhou, et al. 2024. "Coupling of Microbial-Explicit Model and Machine Learning Improves the Prediction and Turnover Process Simulation of Soil Organic Carbon." *Climate Smart Agriculture* 1, no. 1: 100001. https://doi.org/10.1016/j.csag.2024.100001.

Yang, Q., G. Liu, B. F. Giannetti, F. Agostinho, C. M. V. B. Almeida, and M. Casazza. 2020. "Emergy-Based Ecosystem Services Valuation and Classification Management Applied to China's Grasslands." *Ecosystem Services* 42: 101073.

Yu, D., F. Hu, K. Zhang, L. Liu, and D. Li. 2020. "Available Water Capacity and Organic Carbon Storage Profiles in Soils Developed From Dark Brown Soil to Boggy Soil in Changbai Mountains, China." *Soil and Water Research* 16: 11–21.

Yuan, Z., Y. Wang, J. Xu, and Z. Wu. 2021. "Effects of Climatic Factors on the Net Primary Productivity in the Source Region of Yangtze River, China." *Scientific Reports* 11: 1–11.

Zanaga, D., R. Van De Kerchove, W. De Keersmaecker, et al. 2021. *ESA World Cover10 m 2020 v100*. Zenodo. https://doi.org/10.5281/zenodo.5571936.

Zeraatpisheh, M., Y. Garosi, H. R. Owliaie, et al. 2022. "Improving the Spatial Prediction of Soil Organic Carbon Using Environmental Covariates Selection: A Comparison of a Group of Environmental Covariates." *Catena* 208: 105723.

Zhou, Z. H., and J. Feng. 2019. "Deep Forest." *National Science Review* 6, no. 1: 74–86.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.