EEET ECOLOGICAL ENGINEERING & ENVIRONMENTAL TECHNOLOGY

Ecological Engineering & Environmental Technology 2024, 25(11), 200–217
https://doi.org/10.12912/27197050/192397
ISSN 2299–8993, License CC-BY 4.0

Received: 2024.08.09
Accepted: 2024.09.15
Published: 2024.10.01

Evaluation of Machine Learning Models for Predicting Soil Texture Using Sentinel-1A SAR and Topographic Information

Saad Mahmood Sulaiman[1*], Mustafa Ridha Mezaal[1], Alyaa Abbas Ali Al-Attar[1]

[1] Geomatic Engineering Department, Engineering Technical College of Mosul, Northern Technical University, Mosul 41002, Iraq

* Corresponding author's e-mai: suramohammedsami@yahoo.com

ABSTRACT

Applications such as agriculture, hydrology, and environmental management need the mapping of soil texture. In a research region near the Great Zab River in Iraq, this study assessed machine learning models for predicting important soil texture qualities using Sentinel-1A radar and digital elevation data. 75 soil samples in all were gathered, and their percentages of clay, silt, gravel, sand, and moisture content were determined. The models that were examined were artificial neural network (ANN), decision tree (DT), random forest (RF), support vector regression (SVR), and logistic regression (LR). Based on test data, results indicated that RF had the lowest root mean squared error (RMSE) in terms of forecasting clay (0.072 percent), specific gravity (0.011), gravel (10.736 percent), sand (10.213 percent), and silt (1.051 percent). Additionally, it had the greatest coefficient of determination (R2) values for clay (0.900), silt (0.883), sand (0.474), specific gravity (0.519), and gravel (0.568). When it came to predicting moisture content, ANN excelled (RMSE 2.515, R2 0.776). According to the RF feature significance scores, elevation was determined to be the most significant input variable. The study showed that precise maps of soil texture prediction may be obtained by utilizing RF machine learning in conjunction with Sentinel-1A data and digital elevation models. This provides an effective way for mapping soil properties in remote places with minimal effort.

Keywords: soil texture, machine learning, digital elevation model, Mosul City, Al-Zab Area.

INTRODUCTION

Soil texture is a significant land environmental variable that affects a variety of soil processes and qualities. It is defined as the percentage of sand, silt, and clay-sized particles in the mineral component of the soil (Mohammadi et al., 2015). Understanding soil texture is essential for agricultural production, land management, environmental protection, and land use planning, according to Liu et al. (2018), Pacheco et al. (2015), Whisler et al. (2016), and others. Soil texture maps can be helpful in identifying locations with varying soil quality for the purposes of selecting a site, managing crops, and evaluating the influence on the environment (Aliero et al., 2018).

The most accurate technique is to send a sample to a soil testing lab so that the hydrometer method or the pipette method may be used to determine the amounts of sand, silt, and clay. In addition, soil texture may be ascertained in the field by looking at physical soil factors (Vos et al., 2016). Sand is coarse when wet and quickly crumbles if rolled into balls. Although loamy soil is easy to deal with, it has a grainy texture and frequently contains constant ratios of sand, silt, and clay. It may take on the shape of a ball when hydrated, but when squeezed, it crumbles. Silty, dry soil crumbles readily and has the consistency of flour. Pressing wet sand-like soil between fingers and thumb does not produce a ribbon; instead, it feels slick. Surface fissures and big, firm clods are characteristics of clayey soil. Clayey soils feel sticky and are pliable when wet. You may create a ribbon by squeezing wet dirt between your fingers and thumb. More clay is present when a ribbon takes longer to form and breaks. There are several benefits to mapping the amount of sand,

silt, and clay in soil using radar imagery (Gorrab et al., 2015; Ulaby et al., 1996). Radar images have the ability to pierce plants and soil, revealing details about the soil layers under the surface. In remote locations, taking soil samples using conventional methods would be challenging or impossible, this makes it feasible to map the characteristics of the soil. An additional tool for determining soil texture is soil moisture content, which may be estimated using radar scans. By using machine learning algorithms to evaluate radar pictures and extract information about the properties of the soil, detailed soil maps may be created (Zhang and Shi, 2019). There are several restrictions, nevertheless, when it comes to mapping soil texture using radar scans. For example, radar signals may only penetrate a limited depth into the soil; as a result, in regions with thick vegetation or high soil moisture content, they may not fully capture the soil profile. In various terrain, the spatial resolution of radar scans may be insufficient to detect minute variations in soil texture. It takes advanced skills and knowledge to interpret radar pictures and derive soil texture information. This can be difficult and could make the mapping process more unclear. Thus, in addition to radar data, relief factors like topography and landform can improve soil texture mapping (Lu et al., 2017; Mohammed, 2020). According to Laurent et al. (2017), topographic roughness, slope, curvature, local relief, and aspect are some of the terrain characteristics that may have an effect on soil texture mapping, which is then utilized to create geomorphic surfaces. By offering additional details on the topography and environmental factors that affect soil qualities, the combination of these relief features with SAR data might enhance the mapping of soil texture.

LITERATURE REVIEW

Soil texture mapping with laboratory spectra, field-based methodologies, and remote sensing technologies has been extensively researched in the past. Karray et al. (2023) used partial least squares regression (PLSR) and support vector regression (SVR) models to investigate the possible uses of field imaging spectra (IS), laboratory spectra (LS), and their mixtures in the prediction of soil attributes. The principal study subjects were clay, sand, silt, organic matter, nitrate NO^{3-}, and calcium carbonate ($CaCO_3$). SVR exceeded PLSR in predicting soil parameters, with an R2 of 0.79 percent and an RMSE of 1.3 percent. Matazi et al. (2024) used WoSI-ISRIC SoilGrid 250 m data to assess the prediction efficiency of five digital soil mapping (DSM) models. Five models were tested: machine learning (ML)-based models (random forest: RF and random forest residual Kriging: RFRK), Bayesian models, and spatial linear regression (SLR-REML) (Integrated Laplace Approximation-Stochastic Partial Differential Equations: INLA-SPDE and spBAYES). The findings showed that SLR-REML has high accuracy and minimum bias in low spatial autocorrelation circumstances. The two machine learning models that best described nonlinear interactions, RF and RFRK, were the most adaptable when coping with high spatial autocorrelation. The INLA-SPDE model demonstrated flexibility to a variety of data properties. Despite lengthier computation durations, SLR-REML lowered the minimum observation requirement of traditional regression, resulting in better DSM predictions. Furthermore, it was discovered that the most recent version of SoilGrids, with a 250 m resolution, increased the accuracy of global soil data by 60–230 percent. Hengl et al. (2017) attribute the improvement to machine learning, finer resolution covariate layers, and more soil profiles.

Several research have shown that integrating radar data with machine learning algorithms may be used to map soil texture. Machine learning technologies, notably random forest regression, can increase local soil knowledge in West Africa while requiring the least amount of money and labor (Forkuor et al., 2017). The proportions of clay, silt, and sand in the Amazon area were shown to be best predicted by RF, particularly when the P-band of aerial radar was added as a covariate (Ana et al., 2022). A random forest regression (RFR) model was used to generate extremely accurate maps of soil texture and organic carbon in the mid-Himalayas (Rengma et al., 2023). Furthermore, a convolutional autoencoder (CAE) model was presented in individual predictive soil mapping (iPSM) to increase prediction accuracy while decreasing prediction uncertainty for clay, silt, and sand (Liang et al., 2023). Several experiments by Wadoux et al. (2018) indicate that the CNN deep learning model increases the accuracy of mapping soil properties by combining data from several sensors and compensating for measurement errors. Ferreira et al. (2022) mapped the composition of clay, silt, and sand

in a remote area of the Amazon jungle using airborne radar data and machine learning (ML) techniques. The study used airborne radar data to evaluate the prediction capabilities of two sampling procedures, Reference Area, and Total Area, as well as three machine learning (ML) algorithms: regression tree, random forest, and support vector machine. RF gave the most accurate forecasts and incorporating overhead radar P-band data considerably improved prediction accuracy, especially for sand content. Maino (2022) used aerial radiometric surveys and machine learning to determine soil texture in Italy; non-linear models yielded considerable results.

Machine learning paired with fictitious dirt photos may reliably forecast soil qualities in dry regions. This might increase the accuracy of digital soil mapping and lower the cost of soil sampling (Naimi et al., 2021). Synthetic aperture radar accurately recognizes the remaining soil texture groups as clay (35%) and sandy loam (23%). Using RADARSAT-2 polarimetric SAR data as covariates, rather than regular kriging, greatly improves the accuracy of digital mapping for soil surface texture (Niang et al., 2014). Bousbih (2019) used radar and optical data from Sentinel-1 and Sentinel-2 to study soil texture in Tunisia. Using RF, they were able to obtain an overall accuracy of 65%. The soil moisture indicator created by combining Sentinel-1 and Sentinel-2 data yields the best classification results. When Sentinel-1 and Sentinel-2 data are combined with soil moisture indicators, mapping accuracy and texture estimates increase (Bousbih et al., 2019).

These studies demonstrate how useful radar scans and machine learning approaches may be for determining soil texture.

DATA AND MATERIALS

Study area

The research zone stretches 30 kilometers along the Great Zab River, one of the two major tributaries of the Tigris River, from the Kalak to the Al-Gwair districts, east of the Nineveh Governorate (Figure 1). Geology and structural features place the area into the Folded Zone of northern Iraq. Because it may provide water for domestic use, agriculture, and the maintenance of local ecosystems, the Great Zab River is a significant hydrological feature in the area (Osman et al., 2019). Paleozoic volcano-sedimentary rocks inside the structural basement of the Catalan Coastal Ranges' (CCR) horst-and-graben system are the main geological formations of the study region (Ismaiel et al., 2018). Black shales, limestones, and sandstones –all of which may be found in sedimentary and volcanic rock sequences – define these formations. The study region's geography is distinguished by a variety of topographies and elevations. The heights vary from 167 to 465 meters. The topography is mostly level, with an average 2/1000 gradient. Based on elevations, textures, and other characteristics, the research area's hilly topography may be categorized into three groups: high, moderate, and low hills. The topography may have an impact on elements including soil erosion, water movement, and land use patterns. Although there are farmed areas as well, wild colonies make up the majority of the vegetation.

Datasets

The main data sets utilized in this study are the ASTER GDEM and the Sentinel-1A synthetic aperture radar (SAR) image. The 2014 introduction of Sentinel-1A SAR data has a number of uses, including monitoring changes in land cover, agriculture, forestry, and disaster management. Depending on the mode and polarization, Sentinel-1A's SAR imaging has a spatial resolution of 5 to 40 meters, which enables a detailed analysis of the features on Earth's surface. Unlike optical sensors, SAR operates in the microwave region of the electromagnetic spectrum, allowing it to operate day or night and through clouds. Satellite data from Sentinel-1A (S-1A) includes the C-band dual-polarization channels (VV and VH) with a 12-day repeating cycle. Two Sentinel-1A images were obtained for this investigation on January 20, 2024 (Figure 2 and 3).

At 75 different places within the research region, soil samples were taken. Sample locations were chosen along the Greater Zap River at about equal intervals of 400 meters. A 500-gram soil sample was taken from the subsurface of each location (50 cm). The samples were forwarded to the lab for analysis. To test the qualities of the soil, the water content, sieve analysis, and hydrometer analysis were carried out.

NASA and the Japanese Ministry of Economy, Trade, and Industry (METI) collaborated to develop ASTER, the Advanced Spaceborne Thermal Emission and Reflection Radiometer.

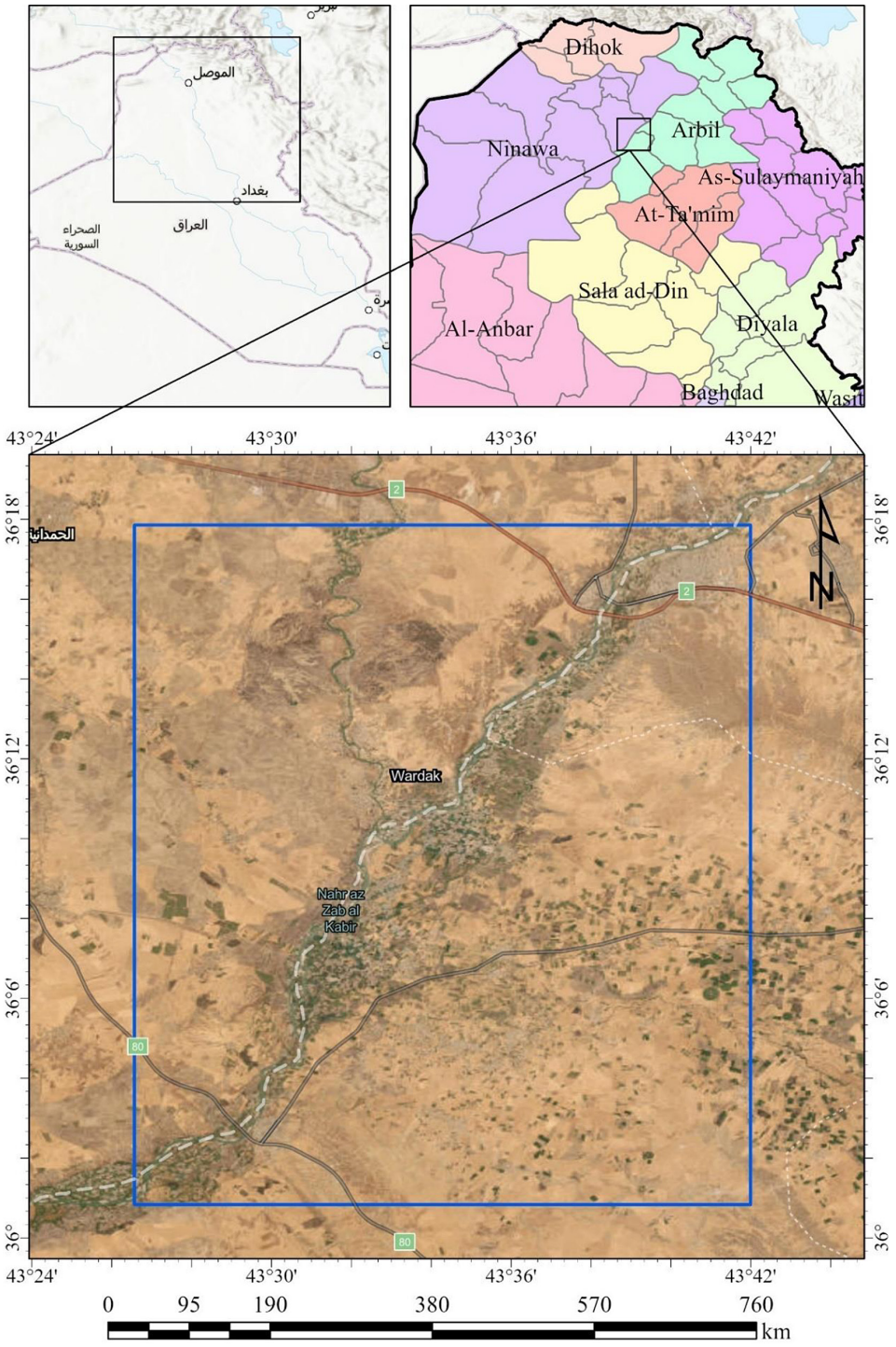

Figure 1. Map of the study area

NASA's Terra probe collected high-resolution topography data for the Earth's surface using stereo-pair photos processed by the ASTER sensor. With a spatial resolution of around 30 meters, the ASTER GDEM provides global coverage and is suitable for a variety of applications such as landform mapping, topography analysis, and natural resource management. Its elevation data is critical to many sectors, including hydrology, urban planning, geology, and environmental studies. The ASTER GDEM V003 data was used for the research region. ASTER GDEM Version 3 maintains the gridding and tile structure of earlier versions, with a spatial resolution of 30 meters and 1° by 1° tiles. The ASTER GDEM Version 3 data product was generated by automatically evaluating the whole ASTER Level 1A library of scenes collected between March 1, 2000, and November 30, 2013.

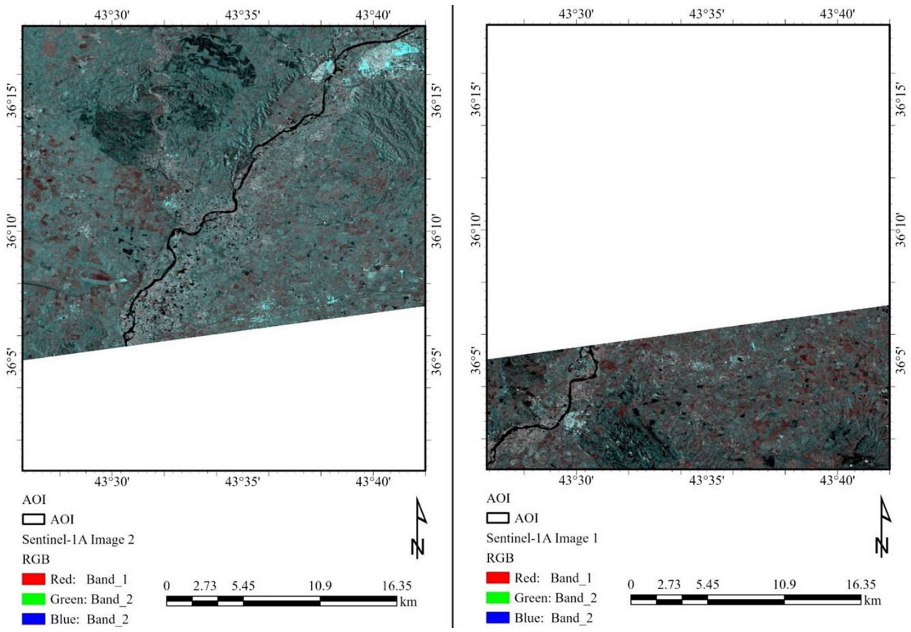

Figure 2. Sentinel-1A image covers the north and south part of the study area

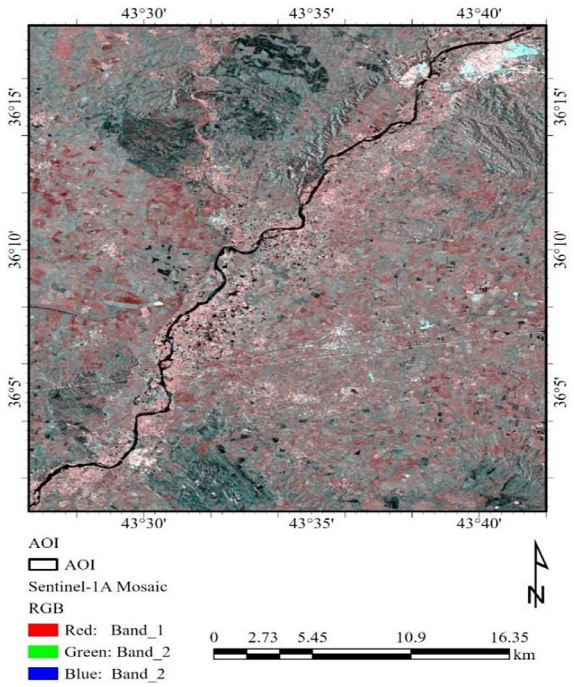

Figure 3. The mosaicked Sentinel-1A image covers the study area

Data preprocessing

Sentinel-1A data was preprocessed using the Bilinear Resampling Method, with terrain correction based on ASTER GDEM data. This phase adjusts for differences in terrain height to ensure geometric precision. The bilinear resampling method was utilized to completely match the radar data to the terrain model. Furthermore, radiometric calibration was employed to get calibrated radar backscatter readings from the sensor's raw digital numbers (DN). Calibrated backscatter values are necessary for quantitative analysis and comparison of radar data across time. Thermal noise alters the radar signal, which can influence the accuracy of later analysis. As a result, any thermal noise injected into the radar data by electrical or environmental variables was eliminated. The linear scale values of the calibrated radar backscatter were converted to decibel (dB) scale. Radar intensity is simple to observe and assess since it is expressed logarithmically on the decibel scale. This change increases the dynamic range of the data and improves the visibility of visual components. Using a Lee Sigma speckle filter decreased the amount of speckle noise seen in the radar image. Speckle noise is a natural component of radar data that may conceal fine details and distort objects. The Lee Sigma filter is a well-known approach for decreasing speckle noise. It uses a statistical technique to maintain picture clarity while smoothing out noise. The 7 × 7 window size specifies the neighborhood in which the filter operates, balancing noise reduction and feature retention.

However, the ASTER GDEM data were georeferenced to achieve precise spatial alignment with other geographic datasets. To allow for integration with other geographic data layers, the data was projected or referenced using the appropriate coordinate reference system (CRS). Furthermore, changes were made to

accommodate for geometric distortions caused by topographical variances. Geocorrection is used to guarantee that topographical characteristics are accurately represented in the elevation model. Filtering and smoothing techniques were used to minimize noise and improve the visual quality of the elevation model.

METHODOLOGY

Overview

The suggested approach for predicting soil texture using relief characteristics, Sentinel-1A data, and machine learning models is shown in Figure 4. Three main datasets were gathered: soil samples from 75 places along the Greater Zap River, which flows through the research region, Sentinel-1A SAR data, and DEM data from ASTER GDEM. Preprocessing of the Sentinel-1A SAR data involved a number of steps, including as radiometric calibration, speckle filtering, dB conversion, thermal noise reduction, and terrain adjustment. Preprocessing methods for the DEM data included geometric rectification and smoothing. The soil samples were prepared in excel sheets identified with unique numbers and included data about moisture content, specific gravity, sieve analysis, and hydrometer analysis. The three datasets after preprocessing were stacked together and stored in a geodatabase. Then, the data was used to perform statistical analyses which included descriptive statistics and correlation assessment. In addition, multicollinearity assessment was used to remove the highly correlated variables that could impact the performance of machine learning models. A final dataset was established for further analysis such as training machine learning models and evaluation of soil texture prediction.

Finally, five machine learning models were trained on the training data and evaluated on the remaining data (test data). The best performed models were used to produce the predicted soil texture maps for the study area (Figure 5).

Relief attributes

Several relief properties were taken from the ASTER DEM data to support the Sentinel-1A

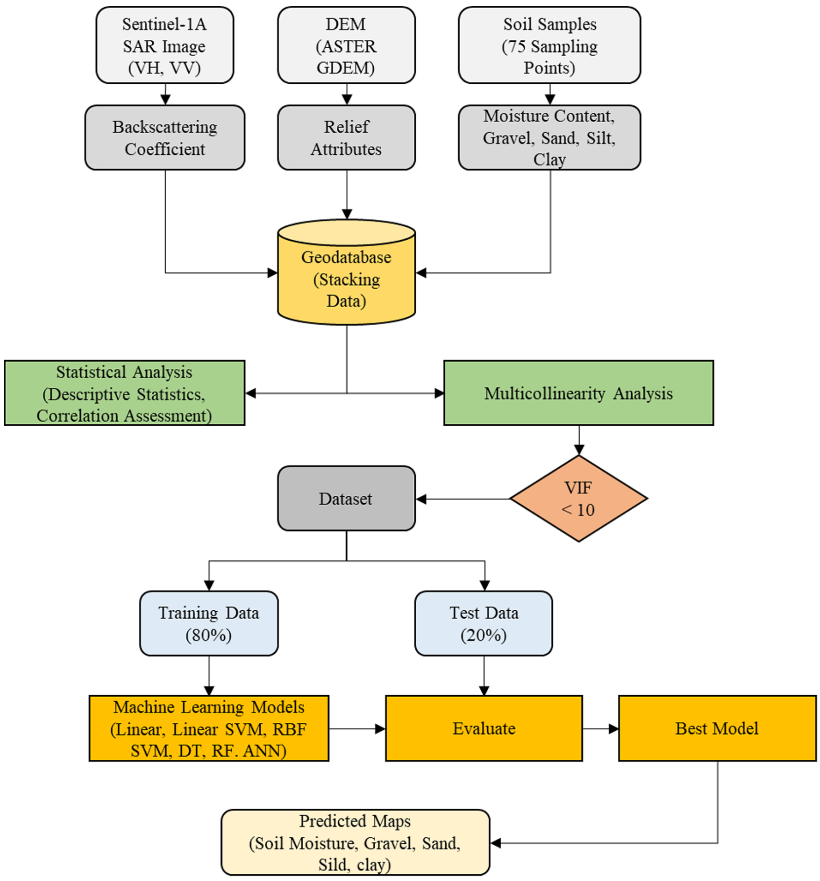

Figure 4. Flowchart of the proposed methodology for soil texture mapping using machine learning algorithms

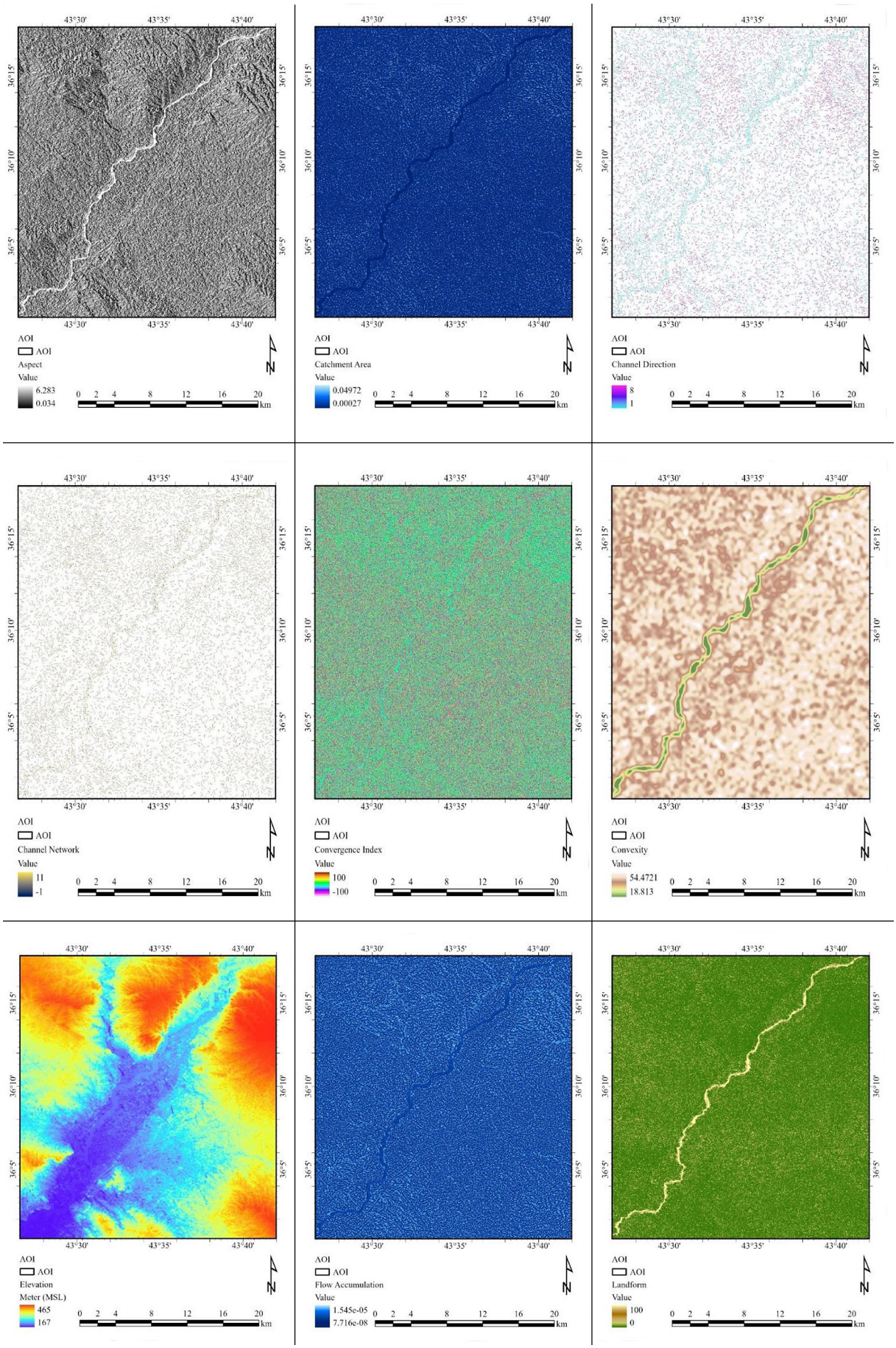

Figure 5. Thematic maps of the input parameters for soil texture prediction

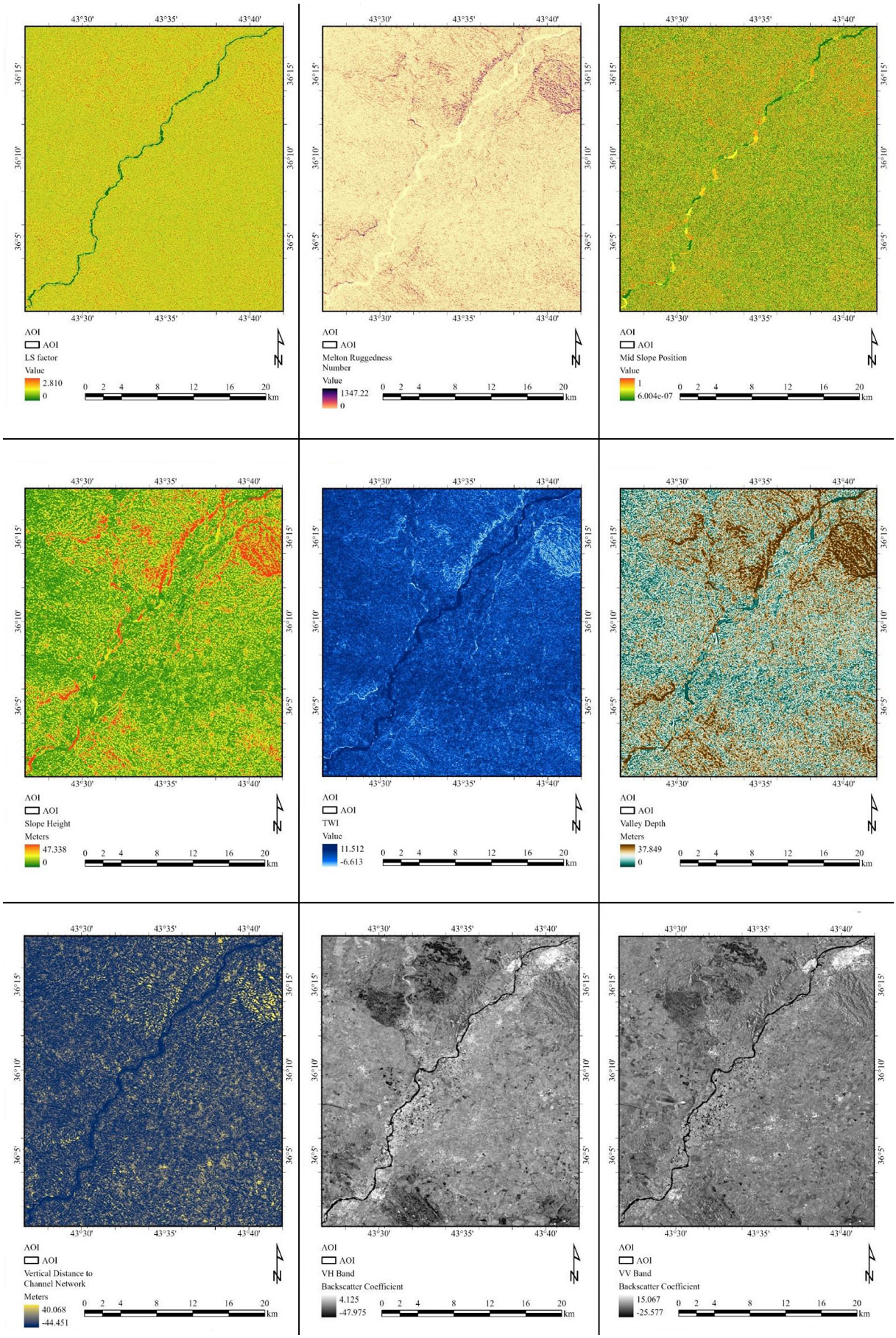

Figure 5. Cont. Thematic maps of the input parameters for soil texture prediction

SAR data, and they were then used to build machine learning models for predicting soil texture. The vertical distance to the channel network, valley depth, topographic wetness index, elevation, landform, flow accumulation, convexity, convergence index, Melton Ruggedness Number, LS Factor, and slope, height, and mid-slope position were among these parameters. Furthermore, six target variables—moisture content, specific gravity, gravel, sand, silt, and clay—were employed to construct the models. Descriptive statistics are shown in Table 1 for the goal variables and relief qualities specified.

Statistical analysis

Two statistical analysis methods were used to perform an initial assessment for the soil texture dataset including descriptive statistics and correlation analysis. The descriptive statistics were used to calculate the basic statistical measures of the Sentinel-1A SAR data bands (VH, VV), relief attributes, and target variables. The calculated statistics included the mean, standard deviation (std.), minimum (min.), quartiles (25%, 50%, 75%) and maximum (max.). The correlation analyses were performed to calculate the correlation among the independent variables and dependent variables using coefficient of determination (R^2).

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(P_i - T_i)^2}{\sum_{i=1}^{m}(P_i - \bar{T})^2} \quad (1)$$

where: T and P are the observed and anticipated values, respectively. The number of samples is denoted by m.

Machine learning algorithms

Linear regression (LR)

A basic statistical method for simulating the connection between a dependent variable and one or more independent variables is called linear regression. The dependent variable and the independent variables are assumed to have a linear connection, meaning that changes in one will correspondingly affect changes in the other.

Mathematically, the basic form of a linear regression model can be represented as:

Table 1. Descriptive statistics of the input data used for machine learning modelling

Variable	Statistic						
	Mean	Std.	Min.	25%	50%	75%	Max.
VH	-18.678	2.154	-23.703	-20.498	-18.612	-16.734	-13.744
VV	-10.637	1.870	-14.837	-11.629	-10.924	-9.543	-3.232
Vertical distance to channel network	2.431	2.127	-3.372	1.229	2.123	2.976	12.324
Valley depth	3.063	2.852	0.058	1.655	2.274	3.590	16.130
Topographic wetness index	-3.449	2.316	-5.423	-4.541	-4.074	-3.625	10.188
Slope	1.501	0.216	0.137	1.570	1.571	1.571	1.571
Slope height	2.599	1.105	1.074	1.688	2.437	3.078	6.974
Mid slope position	0.324	0.173	0.042	0.205	0.297	0.415	0.966
Melton ruggedness number	67.916	80.992	0.000	3.454	37.669	104.750	418.369
LS factor	0.450	0.172	0.031	0.353	0.395	0.498	1.255
Landform	9.778	14.902	0.064	1.492	3.290	12.291	92.909
Flow accumulation	3.3E-07	4.0E-07	1.0E-07	1.0E-07	2.0E-07	4.0E-07	3.3E-06
Convexity	40.692	5.498	32.190	35.518	40.248	44.878	52.163
Convergence index	-0.021	25.808	-59.508	-16.117	0.843	14.708	78.780
Catchment area	8.1E-04	9.5E-04	2.8E-04	3.0E-04	4.5E-04	8.5E-04	7.5E-03
Elevation	222.546	16.648	194.063	208.430	219.836	238.910	251.554
Moisture content	8.416	5.493	2.460	4.275	6.157	12.867	19.904
Specific gravity	2.639	0.016	2.625	2.627	2.630	2.660	2.664
Gravel	43.892	17.169	5.460	32.570	47.630	59.750	65.810
Sand	52.368	15.072	32.810	39.110	48.370	64.550	82.810
Silt	3.562	2.944	0.298	1.674	2.502	5.949	10.909
Clay	0.178	0.220	0.012	0.046	0.121	0.188	0.821

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \quad (2)$$

where: y_i is the i observation of the dependent variable, β_0 is the intercept, representing the value of y when all independent variables are zero, $\beta_1, \beta_2,...,\beta_p$ are the coefficients, indicating the impact of each independent variable x_{i1}, $x_{i2},...,x_{ip}$ on y, $\in_i$ is the error term, accounting for any unexplained variation in y not captured by the model.

The goal of linear regression is to estimate the values of β_0 and β_1 that minimize the sum of squared differences between the observed values of Y and the values predicted by the model. This process is typically done using the method of least squares.

For a multiple linear regression model with p independent variables, the equation becomes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \quad (3)$$

where: X_1, X_2 ,..., Xp are the independent variables. The coefficients $\beta_0, \beta_1, \beta_2,..., \beta_p$ are estimated using techniques like ordinary least squares or gradient descent, aiming to minimize the sum of squared errors.

Linear support vector machine (Linear SVM)

Classification tasks are performed using a supervised machine learning model known as a linear support vector machine (SVM). Unlike linear regression, which predicts a continuous outcome, linear SVM focuses on classifying data points into discrete groups by identifying the hyperplane that divides the classes in the feature space. Choosing this hyperplane optimizes the margin, which is the distance between the hyperplane and the closest data points from each class (also known as support vectors).

Mathematically, the decision function of a linear SVM can be represented as:

$$f(x) = \beta 0 + \beta 1 x1 + \beta 2 x2 + \ldots + \beta p x p \quad (4)$$

where: $f(x)$ is the decision function, x_1, x_2, ..., x_p are the features of the input data point x, $\beta_0, \beta_1, \beta_2, ..., \beta_p$ are the coefficients corresponding to each feature, p is the number of features.

The decision function $f(x)$ classifies a data point x as belonging to one of the two classes based on the sign of $f(x)$. If $f(x)$ is positive, the data point is classified into one class, and if it is negative, the data point is classified into the other class. The hyperplane that separates the classes is determined by finding the optimal values of $\beta_0, \beta_1, \beta_2,..., \beta_p$ such that the margin is maximized. This optimization problem can be formulated as a constrained optimization problem, where the objective is to minimize $\frac{1}{2}\|\beta\|^2$ subject to the constraints:

$$yi(f(xi) - \beta 0 - \beta 1 xi1 - \beta 2 xi2 - \cdots - \\ -\beta p xip) \geq 1 for\ i = 1,2,\ldots,n \quad (5)$$

where: yi is the class label of the ith data point, $x_{i1}, x_{i2},..., x_{ip}$ are the features of the ith data point, n is the total number of data points.

The solution to this optimization problem yields the optimal values of $\beta_0, \beta_1, \beta_2,...,\beta p$, which define the hyperplane that maximizes the margin between the classes.

Radial basis function SVM (RBF SVM)

For applications including regression and classification, the radial basis function (RBF) Support Vector Machine is an effective supervised learning technique. Unlike linear SVMs, which utilize a linear decision boundary, RBF SVMs employ a non-linear decision boundary.

Mathematically, the decision function of an RBF SVM can be expressed as:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b \quad (6)$$

where: $f(x)$is the decision function, x is the input data point to be classified, n is the number of support vectors, a_i are the Lagrange multipliers obtained during training, y_i are the class labels of the support vectors, x_i are the support vectors, $K(x, x_i)$ is the RBF kernel function.

The RBF kernel function $K(x, x_i)$is defined as:

$$K(x, x_i) = exp(-\gamma \|x - x_i\|^2) \quad (7)$$

where: $\|x–x_i\|^2$ is the squared Euclidean distance between x and x_i, y is a hyperparameter that controls the spread of the RBF kernel.

During training, the RBF SVM learns the optimal values of the Lagrange multipliers αi and the bias term b by solving the optimization problem formulated to maximize the margin between the classes while minimizing the classification error. The decision boundary of an RBF SVM is non-linear and can adapt to complex patterns in the data, making it suitable for tasks where the classes are not linearly separable. However, the performance of an RBF SVM heavily depends on the choice of hyperparameters, particularly the regularization parameter C and the RBF kernel parameter y.

Decision tree (DT)

For both classification and regression tasks, a supervised learning approach called a decision tree is employed. This non-parametric model divides the feature space recursively into subsets according to feature values in order to arrive at its conclusions. The ultimate partitions, referred to as leaf nodes, indicate the anticipated result, and each division in the tree corresponds to a decision node.

A decision tree is mathematically expressed as a binary tree structure, with each inner node representing a feature-based judgment and each leaf node indicating the projected conclusion. Each node expresses an opinion based on a certain characteristic's threshold value. During the decision tree building process, each node is assigned the optimum feature and threshold value to enhance information acquisition or decrease impurity. The Gini impurity is a popular measure for determining a node's impurity or purity, and it may be defined as follows:

$$Gini(p) = \sum_{i=1}^{K} p_i(1 - p_i) \quad (8)$$

where: K is the total number of classes and pi is the node's chance of belonging to class I.

When every instance at a node belongs to the same class, the Gini impurity decreases, indicating pure partitions. Entropy is an additional metric that quantifies the degree of uncertainty or unpredictability in a node's class distribution. The following formula is used to find a node t's entropy:

$$Entropy(t) = -\sum_{i=1}^{K} p_{i,t} log_2(p_{i,t}) \quad (9)$$

where: K is the total number of classes and $p_{i,t}$ denotes the percentage of class I occurrences in node t.

By recursively dividing the feature space, the decision tree technique selects the feature and threshold value that maximizes information gain or reduces impurity at each node. The impurity of the parent node is subtracted from the weighted impurity of the child nodes to determine the information gain. Because they are simple to use and intuitive, decision trees are often employed for tasks that need interpretability. However, they are prone to overfitting, particularly in cases when the tree depth is unbounded or noisy data is involved. Pruning, tree depth limitation, and Random Forests are examples of

Random forest (RF)

By combining the decision-making power of several decision trees, Random Forests are a powerful ensemble learning method that improve forecast robustness and accuracy. In order to arrive at the final prediction, it builds several decision trees during training and aggregates their forecasts. Complex mathematical training and prediction methods are used in the process.

First, a large number of bootstrap samples are produced using the random forest method using the initial training dataset. From the training data, a random sampling with replacement is used to create each bootstrap sample. As a result, the data are divided into several subsets, some of which can include duplicates and omit others. An separate decision tree is built for every bootstrap sample. Every node in a tree is constructed by taking into account a random subset of characteristics for possible splits. By introducing variation and keeping the trees from being similar, this randomization helps to reduce overfitting. Forecasts are created by combining the predictions of each decision tree once they have all been built. In classification tasks, the final prediction is chosen by means of a majority vote among all the trees' forecasts. The average of each tree's predictions is usually the final prediction for regression problems. Mathematically, the prediction process can be represented as follows: For regression tasks:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (10)$$

where: $\hat{y}$ is the predicted value, $y_1, y_2, ..., y_n$ are the predicted values from individual decision trees.

Random forests provide various advantages over individual decision trees, including higher prediction accuracy, resistance to overfitting, and flexibility with large and multidimensional datasets. Furthermore, they provide perspectives on the relevance of characteristics, which aid in feature selection and comprehension of the underlying data links. To achieve optimal performance, random forest optimization may need fine-tuning hyperparameters such as the number of trees and the maximum depth of each tree, complicating the training method.

Artificial neural network (ANN)

A multilayer perceptron (MLP) is an artificial neural network composed of many layers of coupled neurons or nodes. It is a popular model for issues like as pattern recognition, regression, and classification due to its strength, adaptability, and ability to mimic complex nonlinear functions. An MLP may be formally defined as a directed

acyclic graph, with numerous neurons connected to neurons in nearby levels inside each layer. To create an output, each layer's neurons use an activation function to process the weighted total of the inputs. The output of one layer becomes the input of the next layer, and so on, until the last layer provides the model's output.

The computation at each neuron in an MLP is explained as follows:

The weighted sum of each neuron j's inputs x_i from the preceding layer $l-1$ is determined as follows:

$$z_j^{(l)} = \sum_{i=1}^{n^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)} + b_i^{(l)} \quad (11)$$

where: $z_j^{(l)}$ represents the weighted sum at neuron j in layer l, $w_{ij}^{(l)}$ represents the weight linking neuron i in layer $l-1$ to neuron j in layer l, $x_i^{(l-1)}$ represents the output of neuron i in layer $l-1$, $bj^{(l)}$ represents the bias term for neuron j in layer l, and $n^{(l-1)}$ represents the number of neurons in layer $l-1$.

The weighted sum $zj(l)$ is processed through an activation function ϕ, introducing nonlinearity into the model and producing the output $aj(l)$ of neuron j in layer l.

$$aj(l) = \phi(zj(l)) \quad (12)$$

Common activation functions include the sigmoid function, hyperbolic tangent function, and rectified linear unit (ReLU) function.

The model's final output is created when the output layer is reached via forward propagation throughout the network. During training, optimization techniques like as gradient descent are used to alter the weights and biases of the MLP in order to minimize a loss function that estimates the difference between the actual and expected outputs.

MLPs may learn sophisticated hierarchical representations of data by including several hidden layers between the input and output layers. MLPs are suitable for a wide range of machine learning applications due to their depth, which enables them to detect complicated patterns and correlations in data. However, difficulties like as overfitting and vanishing gradients make training deep MLPs problematic, necessitating extensive regularization and optimization procedures.

Validation

The proposed prediction models were assessed using three performance metrics. To calculate the metrics for a set of observed (T) and predicted (P) values, use the following formulae. where m is the number of samples. A higher R2 score denotes better agreement between the predicted and actual values. Better prediction performance, however, is indicated by a lower score for the other indicators.

Coefficient of determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(P_i - T_i)^2}{\sum_{i=1}^{m}(P_i - \bar{T})^2} \quad (13)$$

Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m}(P_i - T_i)^2}{m}} \quad (14)$$

Mean absolute error (MAE)

$$MAE = \frac{\sum_{i=1}^{m}|P_i - T_i|}{m} \quad (15)$$

RESULTS AND DISCUSSIONS

Multicollinearity analysis and variable selection results

The VIF values for each variable in the dataset are shown in Table 2 following a multicollinearity analysis. A high VIF led to the removal of six variables: variables such as the topographic wetness index (TWI), slope, LS factor, landform, flow accumulation, and catchment area that have VIF values greater than 10. This suggests that there is a significant problem with collinearity with these variables, which might lead

Table 2. Multicollinearity assessment of the input dataset

Variable	VIF
VH	2.504
VV	2.622
Vertical distance to channel network	4.058
Valley depth	5.858
Topographic wetness index	**133.526**
Slope	**50.256**
Slope height	4.542
Mid slope position	3.021
Melton ruggedness number	4.559
Ls factor	**37.039**
Landform	**48.203**
Flow accumulation	**17.14**
Convexity	4.145
Convergence index	4.263
Catchment area	**58.546**
Elevation	1.551

to problems interpreting the model's regression results. The VIF values of the other variables are moderate: VH, VV, vertical distance to channel network, valley depth, slope height, mid slope position, Melton ruggedness number, convexity, convergence index, and elevation were among the variables that were kept when their VIF was less than 10. These mild VIF scores imply fewer serious issues with collinearity. Eliminating highly collinear variables makes the dataset less redundant and makes it easier to see how each of the remaining variables in the regression model affects each other separately. Although eliminating collinear variables enhances interpretability, it's critical to recognize that doing so may result in the loss of some information.

Table 3 shows that the target variables for gravel, silt, and sand have incredibly high VIF values (9508.239,151.207, and 12360.99, respectively), above the standard cutoff of 10 to indicate multicollinearity. This implies that they have a strong linear connection, which means that the information they offer is quite comparable. Additional variables: the VIF values for moisture content, specific gravity, silt, and clay range from 1.504 to 6.422, all of which are below 10. These numbers imply that there is little to no multicollinearity for these variables.

Performance assessment of machine learning models

The performance evaluations of several machine learning models for predicting soil texture are included in this paper. Moisture content, specific gravity, gravel content, sand content, silt content, and clay content are among the target factors. The assessment criteria used include the R-squared (R^2) score, mean absolute error (MAE), RMSE. On the training set, the ANN model predicts moisture content with the lowest RMSE (2.835) and MAE (2.109). The ANN model has the greatest R^2 value, at 0.736. The RF model has the lowest RMSE of 0.006 and the greatest R^2 of 0.835, the RF model outperformed the others in terms of specific gravity predictions. In the gravel percentage prediction test, the RF model again outperformed the others, with the lowest RMSE of 11.024 and MAE of 7.967. With an RMSE of 10.396 and an MAE of 7.610, the RF model had the fewest errors in predicting sand percentage. The SVR model outperformed the other models in the silt percentage prediction task, with the lowest RMSE of 1.116 and the highest R^2 of 0.911. Finally, the RF model had the highest R^2 of 0.923 and the fewest mistakes in estimating the amount of clay, with an RMSE of 0.060 and MAE of 0.046. The Random Forest model exhibited its robustness in this prediction task by outperforming the bulk of the target variables. While the ANN and SVR models performed less consistently across all targets, they excelled at specialized tasks such as moisture content and silt prediction, respectively. Specific gravity was one of the continuous value predictions that logistic regression found difficult to fit. The ensemble method used by the RF model seems to be successful in teaching it the intricate relationships between inputs and soil texture characteristics. Additional tweaking of hyperparameters and model improvements may enhance performance even more. To sum up, however, the RF model offers a solid foundation for predicting soil texture using this information.

Table 3. Multicollinearity assessment of the target variables

Variable	VIF
Moisture content	2.439
Specific gravity	1.504
Gravel	9508.239
Sand	12360.99
Silt	151.207
Clay	6.422

The test errors for the prediction of moisture content exhibited a pattern akin to that of the training mistakes. Among the models tested, the ANN model had the lowest RMSE of 2.515 and the greatest R^2 of 0.776. On the test data, the RF model's performance somewhat declined. With the exception of ANN, all models had decreases in R^2 between the training and test sets for particular gravity prediction, suggesting a degree of overfitting. Even yet, the RF model managed to attain the best test RMSE of 0.011. Despite somewhat declining from training, the RF model maintained the lowest test RMSE of 10.736 and MAE of 8.089 for the proportion of gravel. Performance drops were less pronounced in the other models. Interestingly, the RF model outperformed the other models and had the best test R^2 of 0.474 on sand percentage prediction. From training to testing, the ANN model's performance significantly decreased. The RF model had a little drop in R^2 to 0.883 on the test set, while the SVR model maintained the lowest test RMSE of 1.364

for silt. The RF model still produced the fewest mistakes on the test set when it came to clay percentage prediction, but its test RMSE of 0.072 and R^2 of 0.900 showed a decline from training. The SVR model performed better during its retention (Table 4).

In conclusion, all models had a little drop in scores from training to testing, but the relative ranks did not change. Overall, the RF and SVR models seemed to be the most resistant to overfitting. The test findings show that RF is highly effective for numerous forecasts of soil texture, while SVR is also effective for some jobs. Additional regularization might aid in enhancing generalization even more.

The significance of the input factors for predicting soil texture is shown in Table 5. With a significance of 1.0, elevation is the most significant

Table 4. Performance assessment of different machine learning models for predicting soil texture targets

Target variable	Dataset	Metric	Model				
			SVR	RF	DT	LR	ANN
Moisture content	Training	Rmse	3.310	2.894	3.125	3.362	2.835
		MAE	1.868	2.114	2.184	2.559	2.109
		R^2	0.641	0.725	0.680	0.629	0.736
	Test	RMSE	3.138	2.635	2.912	3.475	**2.515**
		MAE	1.787	1.999	2.071	2.720	**1.782**
		R^2	0.652	0.755	0.700	0.573	**0.776**
Specific gravity	Training	Rmse	0.017	0.006	0.007	0.010	0.146
		MAE	0.017	0.003	0.004	0.008	0.113
		R^2	-0.144	0.835	0.824	0.585	-87.111
	Test	RMSE	0.017	**0.011**	0.012	0.012	0.175
		MAE	0.017	**0.006**	0.007	0.010	0.134
		R^2	-0.057	**0.519**	0.476	0.496	-114.363
Gravel	Training	RMSE	16.002	11.024	11.390	13.221	16.949
		MAE	10.527	7.967	7.871	10.805	14.454
		R^2	0.147	0.595	0.568	0.418	0.043
	Test	RMSE	14.717	**10.736**	10.802	13.505	16.596
		MAE	10.119	**8.089**	8.400	10.841	13.192
		R^2	0.188	**0.568**	0.562	0.316	-0.033
Sand	Training	RMSE	14.081	10.396	10.405	12.022	17.895
		MAE	9.367	7.610	7.615	10.141	13.803
		R^2	0.151	0.537	0.536	0.381	-0.371
	Test	RMSE	13.115	**10.213**	11.242	11.289	19.941
		MAE	9.074	**8.568**	8.689	8.736	16.198
		R^2	0.133	**0.474**	0.363	0.358	-1.004
Silt	Training	RMSE	1.116	1.742	1.810	2.019	0.864
		MAE	0.700	1.266	1.245	1.604	0.601
		R^2	0.911	0.640	0.611	0.517	0.852
	Test	RMSE	1.364	**1.051**	1.789	2.307	1.738
		MAE	0.921	**0.755**	1.183	1.807	1.205
		R^2	0.803	**0.883**	0.661	0.435	0.680
Clay	Training	RMSE	0.072	0.060	0.061	0.159	0.065
		MAE	0.065	0.046	0.046	0.123	0.051
		R^2	0.891	0.923	0.922	0.465	0.910
	Test	RMSE	0.094	**0.072**	0.095	0.169	0.087
		MAE	0.062	**0.057**	0.057	0.143	0.070
		R^2	0.826	**0.900**	0.825	0.441	0.853

Table 5. Importance of input variables for soil texture prediction

Variable	Importance	Rank
Elevation	1.000	1
Valley depth	0.749	2
VV	0.705	3
Convexity	0.594	4
Slope height	0.487	5
Melton ruggedness number	0.415	6
Vertical distance to channel network	0.301	7
Convergence index	0.141	8
Vh	0.097	9
Mid slope position	0.000	10

variable; valley depth (0.749) and VV are next in importance (0.705). These factors probably reflect important features of the topography, like drainage, erosion, and deposition, that affect the creation of soil texture. The relevance ratings of the remaining factors are smaller, ranging from 0.097 for VH to 0.594 for Convexity. Even while they might not contribute as much individually, they might nonetheless be important in predicting soil texture, particularly when combined with the other top factors. With a 0 significance score, mid-slope position may not be significant for predicting soil texture in this particular situation (Figure 6).

Discussions

The results of the above study are a progression on the trend whereby models that are complex, e.g. random forest and artificial neural network models, give better results than simpler models such as linear regression, in the case of

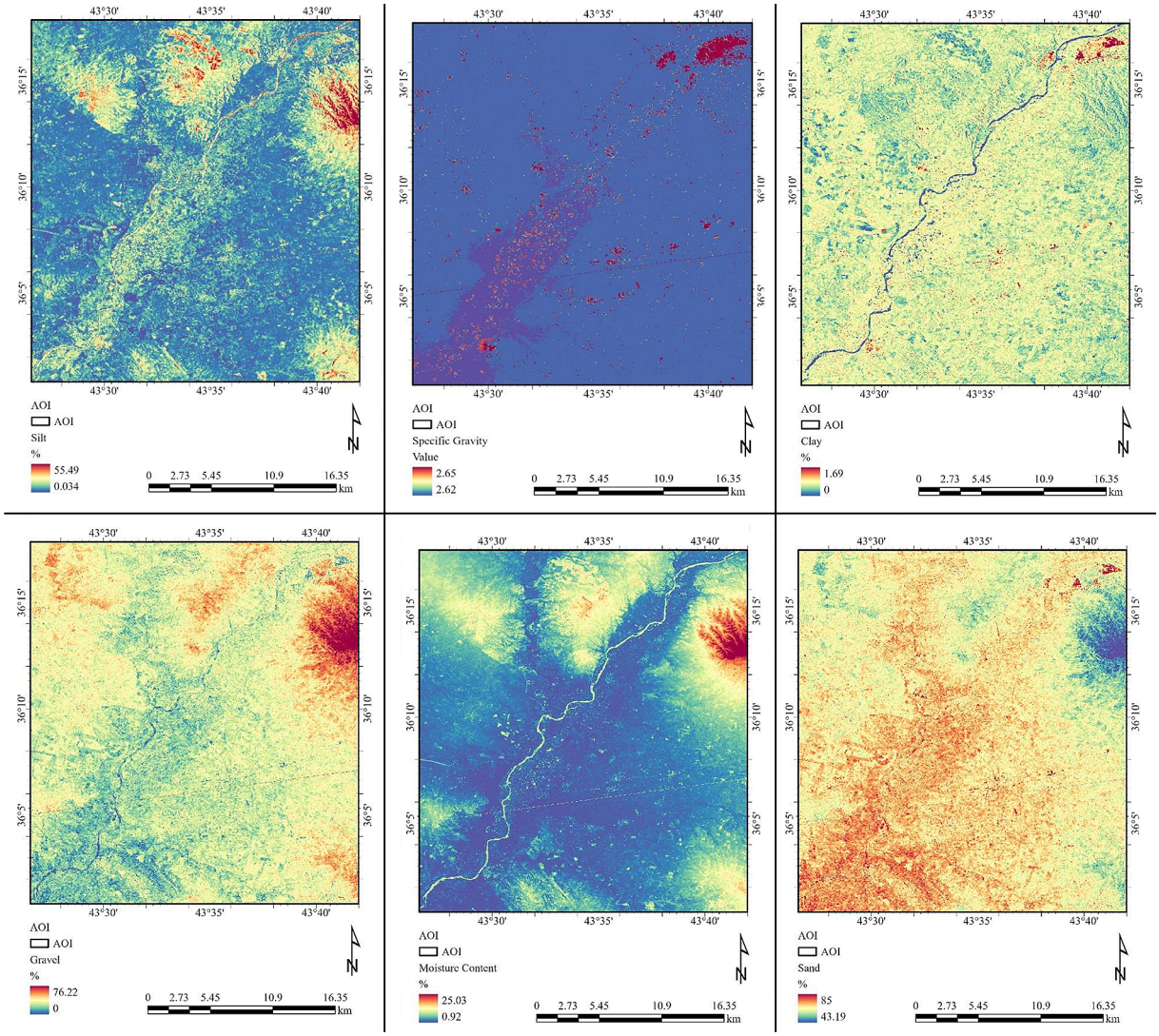

Figure 6. Predicted soil properties maps cover the study area by the proposed machine learning models

soil composition prediction. They do it due to the tendence of complicated models to represent the nonlinear relationships and to detect the small but important details in the soil's structure that are probably used in the classification and prediction tasks.

While all methods remained just above the surface, in most rounds the RF was able to distinguish itself from the field by continuously attributing nearly perfect values to performance measures. This shows that RF is one of powerful models, which accurately predicts soil composition by sensing complicated interactions while preserving the wave of computational efficiency. Ensemble models like random forest proved strong and most successful. Which models are built by accumulation of joint predictions of multiple decision trees rather than the simple addition of such predictions are capable of taking into account the nonlinearity and interactions in the data of soil compositions.

Although the more intricate models that involve processes like decision tree and MLP regress did register much higher accuracy, random forest and artificial neural network models that exude transparency and affordability may be the better choice if this factor is of consequence. The decision trees, particularly, are best fitted for such purposes because they are straightforward and understandable and can be used to spot the main connections that happen between soil features.

Based on the presented results, the random forest, and artificial neural network models are the best choices for predicting soil composition, given their high accuracy, computational efficiency, and potential interpretability. If ensemble-based forecasts are preferable, ensemble models such as random forest can be used, albeit interpretability may suffer.

CONCLUSIONS

This study investigated how Sentinel-1A SAR data and topography information may be used to predict important soil texture qualities using machine learning algorithms. SVR, RF, DT, LR, and ANN were the models that were assessed. The percentages of clay, sand, gravel, silt, and moisture content were the goal variables that were anticipated.

The outcomes show that, in all of the prediction tasks, the RF model performed the best overall. For predicting specific gravity, gravel, sand, silt,and clay percentages on both the training and test sets, it generated the fewest mistakes and the greatest R^2 scores. This illustrates how well RF's ensemble technique captures intricate correlations between the objectives for soil texture and the input variables. Particularly for moisture content prediction, the ANN model did well, while the SVR model was superior at silt percentage prediction. Nevertheless, compared to RF, these models exhibited worse consistency across all targets. The continuous value forecasts were difficult for logistic regression to predict.

The most significant input for predicting soil texture, according to the variable significance analysis, was elevation, which was followed by valley depth and VV backscatter. This is expected as important information about the landscape that influences soil qualities is provided by radar backscatter and topography. Smaller but complementary information was given by the other terrain characteristics.

Overall, the work shows that mapping soil texture, even in remote areas with minimal field sampling, is possible by integrating Sentinel-1A SAR data with digital elevation data and machine learning techniques. Global digital elevation models and publicly available Sentinel-1A images might be used to expand the methodology's application to bigger regions. The models may be able to perform better in generalization with more optimization.

The work is a first step toward creating precise computerized methods for soil mapping. Other remote sensing data sources, such as optical images, should be included into future study to offer more predictor factors. Additionally, more intricate neural network designs may be examined and contrasted with RF. Increasing the sample size might strengthen the robustness of the model. Furthermore, examining the physical explanations for significant variables may provide insightful information on the interplay between soil and landscape. However, this work presents a workable approach to using machine learning and earth observation data for digital soil mapping.

REFERENCES

1. Aliero, M.M., Ismail, M.H., Alias, M.A., Alias Mohd, S., Abdullahi, S., Kalgo, S.H., Kwaido, A.A. 2018. Assessing Soil Physical Properties Variability and Their Impact on Vegetation Using Geospatial Tools

in Kebbi State, Nigeria. IOP Conference Series: Earth and Environmental Science, 169(1), 012111. https://doi.org/10.1088/1755-1315/169/1/012111

2. Ana, C., Ferreira, C., Ceddia, M.B., Costa E.M., Pinheiro, E.F.M., do Nascimento M.M., Vasques G.M. 2022. Use of airborne radar images and machine learning algorithms to map soil clay, silt, and sand contents in remote areas under the Amazon Rainforest. Remote Sensing, 14(22), 5711–5711. https://doi.org/10.1088/10.3390/rs14225711
3. Bousbih, S., Zribi, M., Pelletier, C., Gorrab, A., Lili-Chabaane, Z., Baghdadi, N.N., Aissa, N.B., Mougenot, B. 2019. Soil texture estimation using radar and optical data from Sentinel-1 and Sentinel-2. Remote Sensing, 11, 1520.
4. Bousbih, S., Zribi, M., Pelletier, C., Gorrab, A., Lili-Chabaane, Z., Baghdadi, N., Aissa, N., Mougenot, B. 2019. Soil texture estimation using radar and optical data from Sentinel-1 and Sentinel-2. Remote Sensing, 11, 1520. https://doi.org/10.3390/RS11131520
5. Ferreira, A.C.D.S., Ceddia, M.B., Costa, E.M., Pinheiro, É.F., Nascimento, M.M.D., Vasques, G.M. 2022. Use of airborne radar images and machine learning algorithms to map soil clay, silt, and sand contents in remote areas under the Amazon Rainforest. Remote Sensing, 14(22), 5711.
6. Forkuor, G., Hounkpatin, O., Welp, G., Thiel, M. 2017. High resolution mapping of soil properties using remote sensing variables in South-Western Burkina Faso: A comparison of machine learning and multiple linear regression models. PLoS ONE, 12. https://doi.org/10.1371/journal.pone.0170478
7. Gorrab, A., Zribi, M., Baghdadi, N., Lili Chabaane, Z. 2015. Mapping of bare soil surface parameters from TerraSAR-X radar images over a semi-arid region. In C. M. U. Neale & A. Maltese (Eds.), Remote Sensing for Agriculture, Ecosystems, and Hydrology XVII 9637, 96371F. SPIE. https://doi.org/10.1117/12.2194947
8. Hengl, T., Jesus, J., Heuvelink, G., González, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M., Geng, X., Bauer-Marschallinger, B., Guevara, M., Vargas, R., MacMillan, R., Batjes, N., Leenaars, J., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B. 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE, 12. https://doi.org/10.1371/journal.pone.0169748
9. Ismaiel, I.A., Bird, G., McDonald, M.A., Perkins, W.T., Jones, T.G. 2018. Establishment of background water quality conditions in the Great Zab River catchment: influence of geogenic and anthropogenic controls on developing a baseline for water quality assessment and resource management. Environmental Earth Sciences, 77, 1–12.
10. Karray, E., Elmannai, H., Toumi, E., Gharbia, M.H., Meshoul, S., Aichi, H., Ben Rabah, Z. 2023. Evaluating the potentials of PLSR and SVR models for soil properties prediction using field imaging, laboratory VNIR spectroscopy and their combination. Comput. Model. Eng. Sci, 136, 1399–1425.
11. Laurent, F., Poccard-Chapuis, R., Plassin, S., Pimentel Martinez, G. 2017. Soil texture derived from topography in North-eastern Amazonia. Journal of Maps, 13(2), 109–115. https://doi.org/10.1080/17445647.2016.1266524
12. Liang, S., Zhang M., Wang B. 2023. Predictive soil mapping based on the similarity of environmental covariates using a spatial convolutional autoencoder. Soil Science Society of America Journal, 87(3), 631–643. https://doi.org/10.1002/saj2.20527
13. Liu, J., Huffman, T., Green, M. 2018. Potential impacts of agricultural land use on soil cover in response to bioenergy production in Canada. Land Use Policy, 75(75), 33–42. https://doi.org/10.1016/j.landusepol.2018.03.032
14. Lu, L., Liu, C., Li, X., Ran, Y. 2017. Mapping the soil texture in the heihe river basin based on fuzzy logic and data fusion. Sustainability, 9(7), 1246. https://doi.org/10.3390/su9071246
15. Maino, A., Alberi, M., Anceschi, E., Chiarelli, E., Cicala, L., Colonna, T., De Cesare, M., Guastaldi, E., Lopane, N., Mantovani, F., Marcialis, M., Martini, N., Montuschi, M., Piccioli, S., Raptis, K.G.C., Russo, A., Semenza, F., Strati, V. 2022. Airborne radiometric surveys and machine learning algorithms for revealing soil texture. Remote Sensing, 14(15), 3814.
16. Matazi, A.K., Gognet, E.E., Kakaï, R.G. 2024. Digital soil mapping: a predictive performance assessment of spatial linear regression, Bayesian and ML-based models. Modeling Earth Systems and Environment, 10(1), 595–618.
17. Mohamed, M. 2020. Classification of landforms for digital soil mapping in urban areas using LiDAR Data derived terrain attributes: A case study from Berlin, Germany. Land, 9(9), 319. https://doi.org/10.3390/land9090319
18. Mohammadi, M., Shabanpour, M., Mohammadi, M.H., Davatgar, N. 2017. Characterizing spatial variability of soil textural fractions and fractal parameters derived from particle size distributions. Pedosphere. https://doi.org/10.1016/S1002-0160(17)60425-9
19. Naimi, S., Ayoubi, S., Demattê, J., Zeraatpisheh, M., Amorim, M., Mello, F. 2021. Spatial prediction of soil surface properties in an arid region using synthetic soil image and machine learning. Geocarto International, 37, 8230–8253. https://doi.org/10.1080/10106049.2021.1996639
20. Niang, M., Nolin, M., Jégo, G., Perron, I. 2014. Digital mapping of soil texture using RADARSAT-2 polarimetric synthetic aperture radar data. Soil Science

Society of America Journal, 78, 673–684. https://doi.org/10.2136/SSSAJ2013.07.0307

21. Pacheco, A., McNairn, H., Mahmoodi, A., Champagne, C., Kerr, Y.H. 2015. The impact of national land cover and soils data on SMOS soil moisture retrieval over canadian agricultural landscapes. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(11), 5281–5293. https://doi.org/10.1109/JSTARS.2015.2417832
22. Periasamy, S., Senthil, D., Shanmugam, R. 2019. A soil texture categorization mapping from empirical and semi-empirical modelling of target parameters of synthetic aperture radar. Geocarto International, 36, 581–598. https://doi.org/10.1080/10106049.2019.1618924
23. Rengma, N.S., Yadav, M., Kalambukattu, J.G., Kumar, S. 2023. Machine learning-based digital mapping of soil organic carbon and texture in the mid-Himalayan terrain. Environmental Monitoring and Assessment, 195(8), 994. https://doi.org/10.1007/s10661-023-11608-9
24. Tziolas, N., Tsakiridis, N., Ben-Dor, E., Theocharis, J., Zalidis, G. 2020. Employing a multi-input deep convolutional neural network to derive soil clay content from a synergy of multi-temporal optical and radar imagery data. Remote Sensing, 12(9), 1389.
25. Ulaby, F.T., Dubois, P.C., van Zyl, J. 1996. Radar mapping of surface soil moisture. Journal of Hydrology, 184(1–2), 57–84. https://doi.org/10.1016/0022-1694(95)02968-0
26. Vos, C., Don, A., Prietz, R., Heidkamp, A., Freibauer, A. 2016. Field-based soil-texture estimates could replace laboratory analysis. Geoderma, 267, 215–219. https://doi.org/10.1016/j.geoderma.2015.12.022
27. Wadoux, A., Padarian, J., Minasny, B. 2018. Multi-source data integration for soil mapping using deep learning. SOIL. https://doi.org/10.5194/SOIL-5-107-2019.
28. Whisler, K.M., Rowe, H.I., Dukes, J.S. 2016. Relationships among land use, soil texture, species richness, and soil carbon in Midwestern tallgrass prairie, CRP and crop lands. Agriculture, Ecosystems & Environment, 216, 237–246. https://doi.org/10.1016/j.agee.2015.09.041
29. Zhang, M., Shi, W. 2019. Systematic comparison of five machine-learning methods in classification and interpolation of soil particle size fractions using different transformed data. Hydrology and Earth System Sciences Discussions, 1–39. https://doi.org/10.5194/hess-2018-584