scientific reports

Check for updates

OPEN

Deep learning and data fusion to estimate surface soil moisture from multi-sensor satellite images

Abhilash Singh & Kumar Gaurav ✉

We propose a new architecture based on a fully connected feed-forward Artificial Neural Network (ANN) model to estimate surface soil moisture from satellite images on a large alluvial fan of the Kosi River in the Himalayan Foreland. We have extracted nine different features from Sentinel-1 (dual-polarised radar backscatter), Sentinel-2 (red and near-infrared bands), and Shuttle Radar Topographic Mission (digital elevation model) satellite products by leveraging the linear data fusion and graphical indicators. We performed a feature importance analysis by using the regression ensemble tree approach and also feature sensitivity to evaluate the impact of each feature on the response variable. For training and assessing the model performance, we conducted two field campaigns on the Kosi Fan in December 11–19, 2019 and March 01–06, 2022. We used a calibrated TDR probe to measure surface soil moisture at 224 different locations distributed throughout the fan surface. We used input features to train, validate, and test the performance of the feed-forward ANN model in a 60:10:30 ratio, respectively. We compared the performance of ANN model with ten different machine learning algorithms [i.e., Generalised Regression Neural Network (GRNN), Radial Basis Network (RBN), Exact RBN (ERBN), Gaussian Process Regression (GPR), Support Vector Regression (SVR), Random Forest (RF), Boosting Ensemble Learning (Boosting EL), Recurrent Neural Network (RNN), Binary Decision Tree (BDT), and Automated Machine Learning (AutoML)]. We observed that the ANN model accurately predicts the soil moisture and outperforms all the benchmark algorithms with correlation coefficient (R = 0.80), Root Mean Square Error (RMSE = 0.040 m^3/m^3), and bias = 0.004 m^3/m^3. Finally, for an unbiased and robust conclusion, we performed spatial distribution analysis by creating thirty different sets of training-validation-testing datasets. We observed that the performance remains consistent in all thirty scenarios. The outcomes of this study will foster new and existing applications of soil moisture.

Surface soil moisture is widely used in agriculture, forestry, hydrology, flood and drought prediction, and climate change studies[1]. Depending on the applications, its measurement is required at different spatial (i.e., local, regional, or global), and temporal scales. At the local scale, soil moisture can be measured in the field using dielectric probes/sensors[2], such as; the Time Domain Reflectometry (TDR), Frequency Domain Reflectometry (FDR), and Neutron Probe (NP). These instruments provide highly precise measurements of the surface soil moisture content at different depths. Currently, about 71 International Soil Moisture Networks (ISMN) with more than 2800 operating stations (status as of August 2021) are available worldwide. They provide near real-time point measurements of soil moisture. However, presently the global coverage of the soil moisture network stations are non-uniformly distributed. This causes a data gap, especially in the regions where soil moisture measurement stations do not exist, and if exist, they are sparsely distributed. To overcome these issues, researchers proposed to use satellite images to estimate soil moisture at regional and global scales[2–6]. For example, the European Space Agency (ESA) launched the Soil Moisture and Ocean Salinity (SMOS) mission[7] in November 2009. The SMOS satellite carries an interferometric radiometer that operates at L-band (1.4 GHz). It provides global surface soil moisture at three days revisit time with a spatial resolution of about 30–50 km. Later in January 2015, the National Aeronautics and Space Administration (NASA) launched the Soil Moisture Active Passive (SMAP) under Earth System Science Pathfinder (ESSP) mission[8]. This is equipped with L-band radar sensor and radiometers. It provides daily soil moisture products at a spatial resolution 1–36 km[9]. Despite their global coverage, data voids are present in the SMOS and the SMAP products, particularly at the complex topography, snow-covered, and densely vegetated regions[10,11].

Fluvial Geomorphology and Remote Sensing Laboratory, Department of Earth and Environmental Sciences, Indian Institute of Science Education and Research, Bhopal 462066, Madhya Pradesh, India. ✉email: kgaurav@iiserb.ac.in

To overcome this, researchers explored the potential of Synthetic Aperture Radar (SAR) microwave remote sensing techniques to estimate soil moisture at high spatial and temporal resolutions[12–14]. In the microwave regions, surface soil exhibits a permittivity (ϵ) gradient between the dry and wet soil. For example, the permittivity (ϵ) of dry soil is ≈ 2, and for water, it is ≈ 80[15–17].

Several backscattering models (empirical, semi-empirical, and theoretical) have been proposed to estimate soil moisture from SAR images[18–23]. These models require quad polarised [i.e., Horizontal-Horizontal (HH), Vertical-Vertical (VV), Horizontal-Vertical (HV), and Vertical-Horizontal (VH)] microwave SAR images along with the sensor properties (i.e., wavelength and incidence angle) to estimate soil moisture. One can retrieve soil permittivity and soil roughness parameters from the inversion of the above models. The soil permittivity can be used in Topp's model[24] to obtain soil moisture[25–33]. The quad polarised SAR images are often not available. To overcome this limitation, the existing backscattering models have been modified according to dual polarised SAR images by eliminating one of the unknown parameter (surface roughness) using the in-situ measurements[14,34–38]. The resulting model will have one-equation with one-unknown, which can be solved to get the soil moisture. However, the accuracy of the retrieved soil moisture depends on the accuracy of the surface roughness[39]. These backscattering models assume ideal soil conditions and consider the soil attributes (i.e., soil moisture and roughness) as a stationary process[40]. Such assumptions get violated in the regions where surface topography is complex. Further, these models are validated under control environments at a finer scale. Hence, they may not concede well over a region with significant intra-field variations[41,42].

Recently, machine and deep learning models have emerged as an efficient tool to predict surface soil moisture at high spatial and temporal scales[43–45]. Unlike physical models, machine or deep learning models are data-driven. They combine different relevant input features to map the output. For instance, brightness temperature, SAR backscatter, sensor properties, geographical information, and meteorological variables can be used as input features to setup a machine learning model[4,46,47]. During the training process, a machine learning model learns the soil moisture dynamics solely from the input data. Once the model is developed, its performance can be evaluated from the unseen data. ANN is a widely used machine learning model to estimate soil moisture[48–51].

Ahmad et al.[4] used Tropical Rainfall Measuring Mission (TRMM) and Advanced Very High-Resolution Radiometer (AVHRR) data to estimate soil moisture at 12 km spatial resolution on a daily scale. They trained SVR models at six different sites in their study area by using three input features (i.e., backscatter values, incidence angle from TRMM, and normalised difference vegetation index from AVHRR). For all the sites, they reported that the correlation coefficient ranges from 0.34 to 0.77 and RMSE < 2%. They have also compared the SVR output with ANN and the Multivariate Linear Regression model (MLR). They concluded that the SVR model outperforms ANN and MLR. Santi et al.[50] proposed the ANN-based approach to estimate daily soil moisture at 10 km spatial resolution. They used backscatter, local incidence angle, azimuth angle, Latitude, Longitude information from Advanced Scatterometer (ASCAT), and soil moisture information from International Soil Moisture Network (ISMN) to train the ANN model. They reported that ANN performs well on the testing data sets with R = 0.82 and RMSE = 0.04 m^3/m^3. Lee et al.[43] applied twenty-five different variants of an ANN-based deep learning model to estimate daily soil moisture at 4 km spatial resolution. They used NDVI, outgoing longwave radiation (OLR), solar insolation (INS), broadband albedo (AL), and integrated multi-satellite retrievals for global precipitation measurement (IMERG) as input features to train the models. They observed that the ANN with four hidden layers and 600 neurons in each layer outperforms all other variants with R = 0.89, RMSE = 3.825%, and bias = -0.039. Santi et al.[52] used the ANN model to estimate surface soil moisture from fully polarimetric C-band RADARSAT-2. They used different features from the linear (HV, HH, and VV) and circular polarised data (Stokes vector, RH-RV phase difference, Shannon entropy polarimetric component, alpha angle, RH-RV correlation coefficient, conformity coefficient, and circular polarization ratio).

Adab et al.[53] used a machine learning model to estimate surface soil moisture solely from the optical and thermal images of Landsat-8. They selected Blue (Band-2), Green (Band-3), Red (Band-4), NIR (Band-5), SWIR1 (Band-6), SWIR2 (Band-7), and Land Surface Temperature (LST) as the potential features to train and validate four different machine learning (RF, SVR, ANN, and elastic net) algorithms. They concluded that the RF is able to predict surface soil moisture accurately. The final soil moisture product has a spatial resolution of 30 m and a temporal resolution same as the Landsat-8 (i.e., 16 days). Datta et al.[54] proposed a regression-based machine learning model to estimate surface soil moisture from Sentinel-1 images at 12 days temporal resolution. They used VV, and VH polarised images as the input features to train the models (RF, SVR, linear regression, multiple linear regression, and K-nearest neighbors). They used 40 and 16 in-situ samples to train and validate the models, respectively. They observed RF outperforms all the other models with R = 0.93 and RMSE = 0.03 m^3/m^3. Greifeneder et al.[55] applied the gradient-boosted regression tree-based approach on the Google Earth Engine platform to estimate surface soil moisture at a 50 m spatial scale at every 12 days. They used soil moisture data from ISMN and features from Landsat-8 and Sentinel-1 to train the machine learning model. They reported that the model performs reasonably well on the test data with R = 0.90 and RMSE = 0.04 m^3/m^3.

The studies discussed above suggest a trade-off between the spatial and temporal resolutions of the soil moisture products. High temporal resolution soil moisture products are available at coarser spatial resolution and vice-versa. This depends on the selection of input features and soil moisture products used to train the machine learning model. Training machine learning models with the in-situ measured soil moisture by utilising the potential of multi-sensor Sentinel-1 (A & B) concurrently with multi-sensor Sentinel-2 (A & B) can provide soil moisture product at optimal spatial and temporal resolutions. This study aims to improve the spatial (60 m) and temporal (6 days) resolutions of the soil moisture product through dual polarised Sentinel-1 backscatter, red and near-infrared reflectance from Sentinel-2, and SRTM elevation data by using data fusion and a deep learning approach. We propose a novel architecture based on a fully connected feed-forward ANN to estimate surface soil moisture at high spatial and temporal resolutions from microwave (Sentinel-1), optical (Sentinel-2), and topographic (SRTM-DEM) data. For training the ANN architecture, initially, we have selected

seven satellite-derived features as input predictors. We have also generated two new synthetic features through linear data fusion of existing features. We used the in-situ measured surface soil moisture as a response variable. We trained the model by using the input features as predictors and in-situ measured surface soil moisture as response. Finally, we compared the performance of our model with the benchmark algorithms (GRNN, RBN, ERBN, GPR, SVR, RF, Boosting EL, RNN, BDT, and AutoML) and selected the best model to estimate surface soil moisture on the Kosi Fan.

Study area

We conducted this study on the Kosi Fan in the Himalayan Foreland in north Bihar plain, India (Fig. 1). It is one of the largest fluvial fan, spread over an area of about 10, 351 km^2. This is a result of frequent avulsions of the Kosi River channels. In the last three centuries, the Kosi River has migrated about 150 km[56–59]. During the process of migration, the river has deposited its sediments and built a large fan like structure. This fan has been active since the Holocene[60]. The Kosi Fan surface is composed of homogeneous quartz grains with a median size varying from medium sand (300 μm) to fine sand (100 μm) near its apex and toe, respectively[59,61]. The topography of the Kosi Fan is nearly flat with a small gradient from 8×10^{-4} in the proximal to 6×10^{-5} in the distal part[59]. The Kosi Fan falls in a tropical climatic zone. The minimum and maximum average annual temperature vary in a range between 18 to 32 °C. The temperature is maximum during the summer and approaches minimum during the winter (Fig. 2). The relative humidity ranges between 30 and 90%. It is maximum (70–90%) during the monsoon (i.e., June–September) and minimum (30%) in the early summer (i.e., March–April). On an average, the Kosi Fan receives about 1484 mm rainfall annually[62,63]. Most of the rainfall (about 80%) occurs during the Indian Summer Monsoon (June–September). The groundwater on the Kosi Fan remains shallow throughout. It varies from 1.8 to 8.1 m and 1.0 to 6.4 m below ground level (bgl) during the pre-and post-monsoon periods, respectively (http://cgwb.gov.in/). Apart from this, one can see numerous waterlogged patches and isolated channels throughout the Kosi Fan. Most part of the Kosi Fan gets flooded every year during the monsoon period. The dominant landuse and landcover types are agricultural (≈ 84%), surface water bodies (≈ 9%), and built-up areas (≈ 7%)[64].
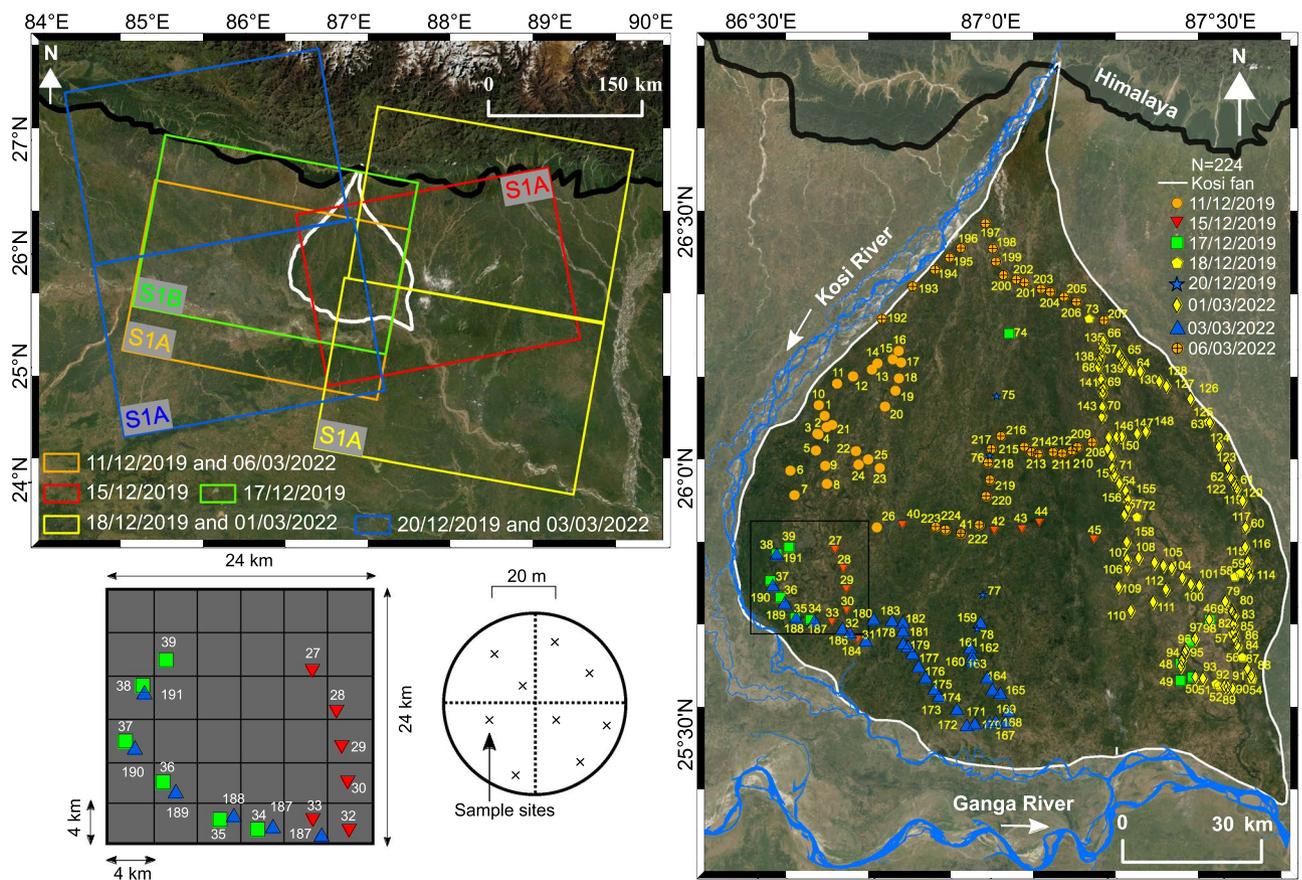
Figure 1. Image on the top left panel shows the rectangular footprints of the Sentinel-1 satellite images of different dates. The boundaries of the footprints are available at the European Space Agency Copernicus Open Access Hub website (https://scihub.copernicus.eu/). Image on the right panel shows the in-situ measured locations with sample ID (in different shapes) and their acquisition dates. The square grid on the bottom left panel illustrates the random sampling strategy. The maps are created in ArcGIS v10.8 software available at Environmental Systems Research Institute (ESRI) website (https://www.esri.com/en/arcgis/about-arcgis/overview).
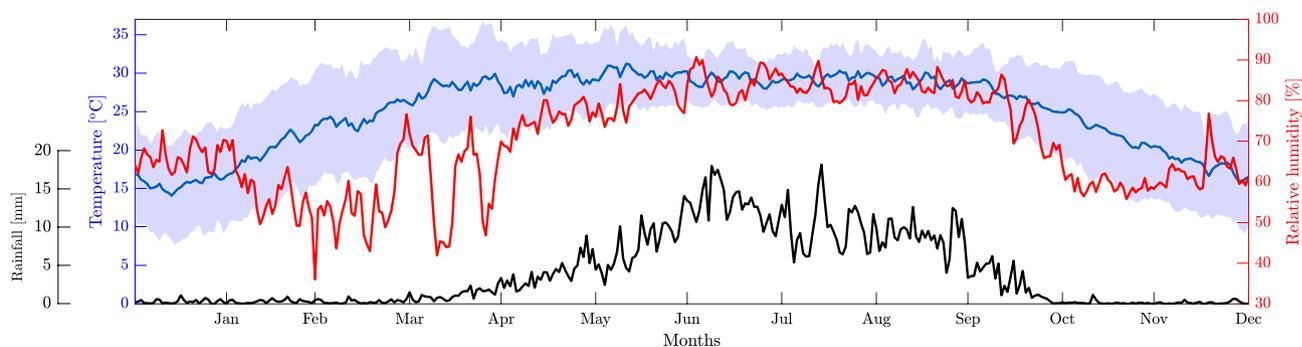
Figure 2. Time series plot (daily average) of the rainfall, temperature, and relative humidity over the study area from 1980 to 2018. The limits of shades in blue represent the maximum and minimum (temperature) values.

Altogether, the Kosi Fan is an ideal field site to study soil moisture variability. Such knowledge can be very useful for better planning of agriculture, flood, drainage congestion, and waterlogging predictions on the Kosi Fan.

Datasets

Satellite. We use publicly available Sentinel-1 (SAR), Sentinel-2 (optical) images, and digital elevation model (DEM) from the shuttle radar topographic mission (SRTM). We downloaded the Sentinel images from official website of the European Space Agency (https://scihub.copernicus.eu/) and SRTM-DEM from the Geological Survey (USGS) website (https://earthexplorer.usgs.gov). Table 1 reports the detailed descriptions of the Sentinel images.

European Space Agency (ESA) launched the Sentinel-1A (on 3rd April 2014) and Sentinel-1B missions (on 25th April 2016) as two-satellite constellation under the Copernicus Programme (formerly known as Global Monitoring for Environment and Security). The revisit time of these satellites is 12 days. However, when considered together, a revisit time of six days can be achieved[14,65]. Sentinel-1A and 1B operates at a frequency of 5.405 GHz. At this frequency, microwave signals can penetrate up to 5 cm below the dry soil column[66–68].

The Sentinel-1A and 1B acquire images in strip map mode, interferometric wide swath mode, extra-wide swath mode, and wave mode. Depending upon the acquisition mode, the SAR products are available at three levels; level-0 (unfocused SAR raw data), level-1 (Single Look Complex (SLC) and Ground Range Detected (GRD) data), and level-2 (Ocean geophysical product derived from level-1). For wave mode, only single polarisation is available i.e., either VV or HH. For the remaining modes, dual polarised images are available i.e., either VV+VH or HH+HV. For the polar environment, sea-ice zone HH or HH+HV polarised data is available. The VV or VV+VH polarised data is available for all other observation zones at 10 m × 10 m cell size with 250 km swath. We have used VV+VH dual polarised GRD (level-1) images in this study.

In continuation of Sentinel missions, the ESA launched Sentinel-2A (on 23rd June 2015) and Sentinel-2B (on 7th March 2017) as a constellation of two polar-orbiting satellites. They provide data in two levels; level-1C (top-of-atmosphere corrected), and level-2A (bottom-of-atmosphere corrected)[69]. Sentinel-2 (A & B) together have a revisit time of 5 days[70]. They acquire images of the earth in 13 different spectral bands from Visible Near

Sentinel-1				
Date (dd/mm/yyyy)	Polarisation	Incidence angle (°)	Pixel size (m × m)	Pass
11/12/2019	Dual (VH, VV)	38.6	10 × 10	Descending
15/12/2019	Dual (VH, VV)	38.5	10 × 10	Ascending
17/12/2019	Dual (VH, VV)	38.4	10 × 10	Descending
18/12/2019	Dual (VH, VV)	38.5	10 × 10	Descending
20/12/2019	Dual (VH, VV)	38.5	10 × 10	Ascending
01/03/2022	Dual (VH, VV)	38.4	10 × 10	Descending
03/03/2022	Dual (VH, VV)	38.5	10 × 10	Ascending
06/03/2022	Dual (VH, VV)	38.6	10 × 10	Descending
Sentinel-2				
Date (dd/mm/yyyy)	Orbit number and direction	Bands	Wavelength (nm)	Spatial Resolution (m)
09/12/2019	76, Descending	4, 8	646–685, 774–907	10
03/03/2022	76, Descending	4, 8	646–685, 774–907	10

Table 1. Detailed descriptions of the Sentinel-1/2 images.

Infra-Red (VNIR) to Short Wave Infra-Red (SWIR) of the electromagnetic spectrum. For our purposes, we have used band-4 (Red; 665 nm) and band-8 (NIR; 865 nm). These bands have a spatial resolution of 10 m.

In a joint venture with National Geospatial-Intelligence Agency (DoD/NGA), the German Aerospace Center (DLR), and Agenzia Spaziale Italiana (ASI), NASA launched an 11 days SRTM shuttle mission in February 2000. It contained two independent SAR sensors in C-band (NASA) and X-band (DLR/ASI). Currently, three versions of SRTM are publicly available. This includes SRTM non-void filled, SRTM void filled, and SRTM 1 arc-second global[71]. We downloaded the void-filled DEM of spatial resolution 1 arc-second (30 m).

Field measurement. In the field campaigns during December 11–20, 2019 and March 01–06, 2022, we measured soil moisture on the Kosi Fan by using the ML3 theta probe (Fig. 3a). We calibrated the theta probe by using the procedure explained in Singh et al.[14]. We adopted universal random grid sampling method to measure soil moisture in the field. We divided the study area into small square grids of 4 km × 4 km (Fig. 1) and randomly selected the grids for measurements. To ensure the same moisture content is illuminated and recorded by the Sentinel-1 SAR pulses, we measured the surface soil moisture at 5 cm depth from the topsoil layer[66–68]. To minimise the spatial heterogeneity, we collect 7–10 in-situ measurements over the footprint of the satellite pixel (i.e., 60 m). Each measurement is separated by at least 20 m. Finally, we calculate the average value of these measurements to get a representative value of soil moisture in a grid. This enables us to perform the direct point-to-pixel comparison of soil moisture[72,73]. We collected 224 such measurements over the entire study area from the apex to the toe of the Kosi Fan. At each sampling location, we have also recorded their coordinates using the Garmin GPSmap-64s (Fig. 3b). Figure 3c illustrates the ground conditions of some of the sampling sites (#10, #19, #52, #69, #71, #82, #109, #119, #129, #145, #153, #183, #193, and #215).

Methodology

Figure 4 illustrates the detailed methodology adopted in this study. Firstly, we process the satellite images to extract input features to be used in the machine learning model. Secondly, we perform the feature engineering and setup a feed-forward multi-layer ANN model for training, validation, and testing. Finally, we evaluate the performance of ANN in terms of error analysis, computational time complexity analysis, and spatial distribution
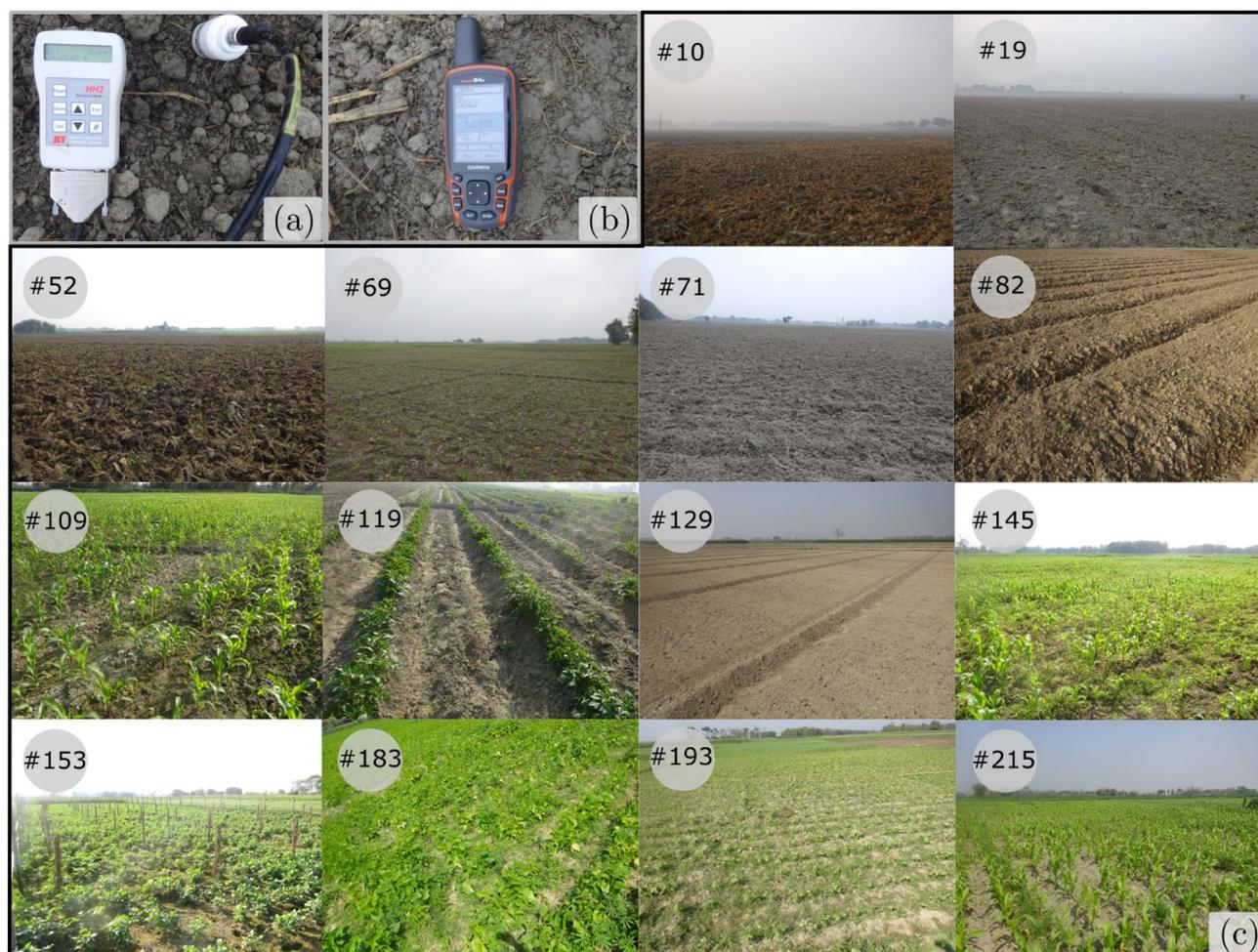
Figure 3. Field photographs illustrate the ground conditions (#10, #19, #52, #69, #71, #82, #109, #119, #129, #145, #153, #183, #193, #215) of the study area at the time of soil-moisture measurement .

analysis and compare the output of the ANN model with ten different benchmark algorithms (i.e., GRNN, RBN, Exact RBN, GPR, SVR, RF, Boosting EL, RNN, BDT, and AutoML). The detailed feature extraction and model setup steps are discussed in the following subsections.

Feature extraction. *Image processing.* We used the Sentinel Application Platform (SNAP v8.0) to process Sentinel-1 images. It is an open-source Earth Observation processing tool. We performed the radiometric calibration, multi-look correction (with a multi-looking factor of 6), speckle noise removal, and terrain correction to process the raw Sentinel-1 images. The resulting backscatter (σ_0) images for both polarisations (VV and VH) have the grid size of 60×60 m. This is because of the multi-looking process, which averages the adjacent oblong pixels to a square pixel of size 60 m (raw pixel size multiplied by the multi-looking factor). We also processed the Sentinel-2 images to compute the normalised vegetation index (NDVI). We do this by taking a ratio of the difference between near-infra-red and red bands to their sum. The resulting NDVI image has a spatial resolution 10×10 m. The NDVI image has pixel values between -1 to $+1$. The higher values of NDVI represent healthy vegetation[74].

Feature selection and scaling. The prediction accuracy of any machine learning model primarily depends on the quality of input data. Without high-quality datasets, even high-performing machine learning algorithms are rendered ineffectual. In addition, the data pre-processing task is required to transform the raw data into a form that is more suitable to the machine learning model, which increases the efficiency and accuracy of the model. We have performed three feature engineering operations (i.e., feature extraction, generation, and scaling). Initially, we have extracted seven features from Sentinel-1, Sentinel-2 images, and DEM data. These are backscatter values (σ_0) in VV and VH polarisations, and incidence angles from Sentinel-1 images. The radar backscatters (VH and VV) are highly sensitive to soil moisture due to the presence of dielectric gradient[16]. The SAR incidence angle is an important sensor parameter that influences satellite-derived soil moisture[75,76]. We obtained the NDVI values from Sentinel-2 and elevation from the mean sea level at each pixel from the DEM. We also obtained the coordinates (Latitude and Longitude) of each pixel of the input images. Vegetation descriptors such as NDVI are important to incorporate the impact of vegetation on soil moisture retrieval. The dependency of soil moisture on the surface elevation is well known and frequently used in observing the soil moisture pattern and machine learning models[77–80]. To incorporate the spatial dependencies of data in spatial machine learning applications, we included geolocation (i.e., Latitude and Longitude) variables[50,81,82]. Further, we have also generated two synthetic features (i.e., VH/VV and VH-VV) from the existing feature set (i.e., VH and VV) by using the linear data fusion of VH and VV[39]. These synthetic features are more sensitive towards the dielectric and geometric properties of soil[76,83,84].

We now use the nearest neighbour resampling method to resample the image pixels of the input features to a common grid size (60×60 m). Finally, we apply the standard z-score scaling to scale all the nine features at the same level according to Eq. (1);

$$f_{scale} = \frac{f_{initial} - \overline{f}}{f_{std}} \tag{1}$$

where $f_{initial}$ represent the initial feature, $\overline{f}$ represent the mean value, and f_{std} is the standard deviation of the feature.
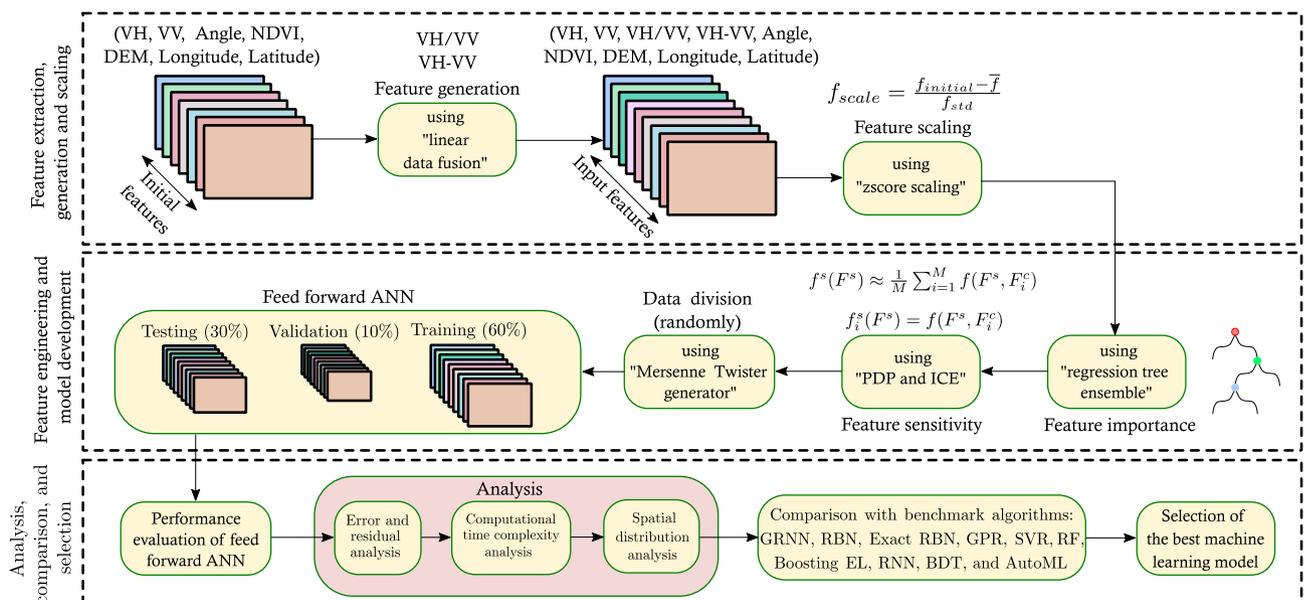
Figure 4. Flowchart illustrates the detailed methodology.

Feature importance and association. We evaluate the relevancy of our input features in predicting the response variable (i.e., soil moisture). We compute the importance score of individual features by using the regression ensemble tree approach. We boosted five hundred regression trees (i.e., $m = 500$), each with an unity learning rate (i.e., $\gamma=1$). We do this by using the Least Squares gradient Boosting (LSBoost) algorithm. We have considered the traditional decision tree as a weak learner (i.e., decision stumps). The LSBoost algorithm starts training a single weak learner at a time and simultaneously identifies its weak points. Based on these weak points, it creates a new weak learner (h_i) and computes the corresponding weight (i.e., δ_i). Finally, the current model (L_i) is updated by the algorithm by focusing on the weak point of the previous weak learner (L_{i-1}) according to Eq. (2);

$$L_i = L_{i-1} + \gamma \cdot \delta_i \cdot h_i \qquad (i = 1, 2, 3, \ldots, m) \tag{2}$$

After training, it incorporates the weak learner into the current model. It then iteratively generates the ensemble of weaker learners (i.e., a single strong learner, L_m). Now, we estimate the entire changes in node risk that result from splitting on each feature, normalising it in relation to the total number of branch nodes (i.e., NR_{branch}), and using that information, we compute the relative feature relevance score. Mathematically, the changes in the node risk (i.e., ΔNR) is computed according to Eq. (3);

$$\Delta NR = \frac{NR_p - (NR_{c1} + NR_{c2})}{NR_{branch}} \tag{3}$$

where NR_p represents the node risk of the parent node and NR_{c1} & NR_{c2} represents the node risk of two children. The node risk at individual node (NR_i) is calculated according to Eq. (4);

$$NR_i = P_i \cdot MSE_i \tag{4}$$

where P_i and MSE_i represents the node probability and mean square error of node i, respectively.

Further, to measure the correlation amongst the features, we calculate the feature association matrix (9×9 matrix). The existence of any correlated features adversely affects the machine learning model by making the model unstable and more sensitive to uncertainty. The values in the matrix represent the similarities between the decision rule that split on each observation. A higher value for a pair of features suggests that they are highly correlated.

Feature sensitivity. The feature importance score only suggests the relevancy of a feature with respect to the response variable (*i.e.,* soil moisture). To understand the association (either positive or negative) between the features and response variable, we need to analyse the Partial Dependency Plot (PDP), and Individual Conditional Expectation (ICE) curves[85,86]. PDP explains the partial dependency of each feature of the feature data (*i.e.,* $F = \{f_1, f_2, \ldots, f_n\}$, where n is the total number of features) on soil moisture by marginalising the impact of all other features. Whereas the ICE is an advancement of PDP that evaluates the feature impact on soil moisture for each observation. We created a subset $F^s = \{f_1\}$ and a complimentary set F^c of F^s, any prediction on F can be computed according to Eq. (5);

$$f(F) = f\left(F^s, F^c\right) \tag{5}$$

We can now estimate the partial dependence of the feature in F^s by computing the expectation (E_c) of Eq. (5);

$$f^s(F^s) = E_c\left[f\left(F^s, F^c\right)\right] \tag{6}$$

$$= \int f\left(F^s, F^c\right) \cdot mp_c(F^c) \cdot dF^c \tag{7}$$

where $mp_c(F^c)$ is the marginal probability of F^c according to Eq. (8);

$$mp_c(F^c) \approx \int p\left(F^s, F^c\right) \cdot dF^s \tag{8}$$

Finally, we can compute the partial dependency of the feature in F^s according to Eq. (9);

$$f^s\left(F^s\right) \approx \frac{1}{M} \sum_{i=1}^{M} f\left(F^s, F_i^c\right) \tag{9}$$

where M is the total observations. Finally, we disaggregate Eq. (9) to obtain the ICE curves according to Eq. (10).

$$f_i^s\left(F^s\right) = f\left(F^s, F_i^c\right) \tag{10}$$

Model setup.

Feed-forward ANN. In a feed-forward ANN model, the basic computations performed by each neuron are used to predict the model performance. This is a two-step process. In the first step, individual inputs of the neuron (i.e., input vector, x) and the corresponding weight values (i.e., weight vector, w) are combined together by a summation function. The output of a summation function is a dot product of weight vectors and input vectors (i.e., $w \cdot x$). A bias (or threshold) is added to the dot product forming the output (f) according

to Eq. (11). In the second step, output (*f*) is fed into the argument of an activation (or transfer) function, which is then used to calculate a scalar value.

$$f = \sum \boldsymbol{w} \cdot \boldsymbol{x} + b \tag{11}$$

In multi-layer feed-forward ANNs, the network architecture consists of *N* neurons that are organised in *L* layers (*L* > 1). The first layer is input layer, which comprises input variables. Each neuron in a layer *l* (1 ≤ *l* ≤ L) is connected to all the neurons present in the previous layer (*l* − 1) or to the input layer if; *l* = 1 (Fig. 5). This way, the computation (or information) flows from the first to the last layer (forward propagation). The output layer (*l* = *L*) is the last layer of the neuron, whereas the other layers of the neuron are referred to as hidden layers. This study uses three hidden layers, each with five neurons. The neuron input is constructed as a linear combination of its received input values that correspond to the output of the previous layers[87] according to Eq. (12);

$$a^l = \sum \boldsymbol{w}^l \cdot \boldsymbol{x}^{l-1} + b^l \tag{12}$$

where, a^l is the input of a neuron present at *l* layer, $\boldsymbol{w}^l$ is the weight vector for the neurons present at *l* layer, $\boldsymbol{x}^{l-1}$ is the output of a neuron present at *l* − 1 layer, and b^l is the bias value at layer *l* which is followed by an activation function.

Activation function. The choice of an activation function determines how a network maps the input features to the output[88,89]. Different layers can have different activation functions, and their selection can strongly influence the performance of a feed-forward ANN in terms of complexity and accuracy. However, there is no universal rule to select the activation functions. The identity (or linear) activation function (purelin) is almost always used at the input and output layers, whereas non-linear activation functions are generally preferred at the hidden layers. The most commonly used non-linear activation functions are hyperbolic tangent sigmoid (tansig) and log-sigmoid (log-sid)[90,91].
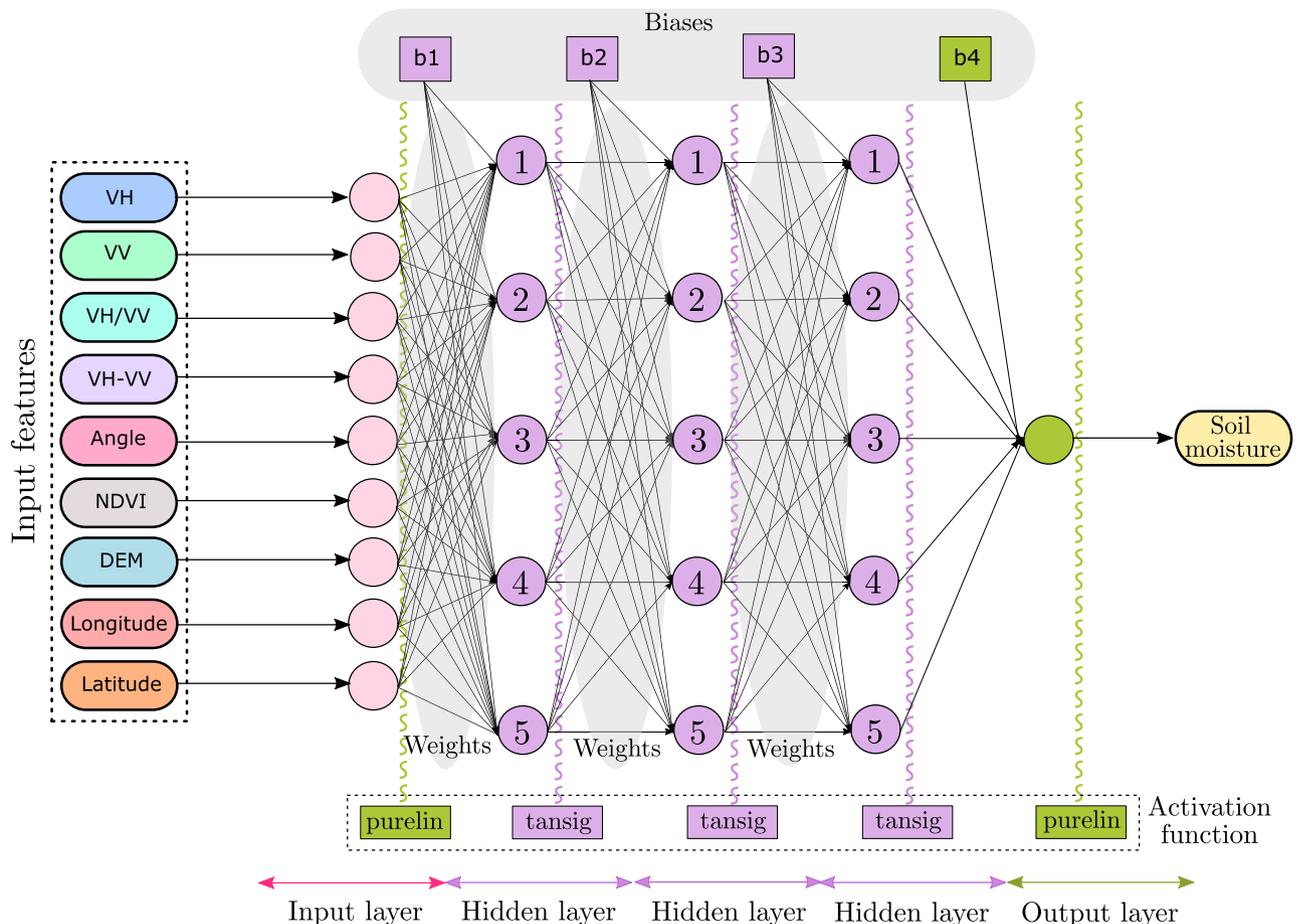
Figure 5. A fully connected 9-5-5-5-1 (I-H_1-H_2-H_3-O) feed-forward ANN architecture for soil moisture estimation. I, H, and O represent the input, hidden, and output layers, respectively.

We have used the identity activation function for the input and output layers and the hyperbolic tangent sigmoid activation function for all the hidden layers. Mathematically, the hyperbolic tangent sigmoid is expressed as Eq. (13);

$$T = \frac{2}{\left(1 + e^{-2 \cdot n}\right) - 1} \tag{13}$$

This is mathematically analogous to tanh(n). We are using Eq. (13) instead of tanh(n) due to its lower time complexity. To reduce the computational time complexity, usually, the fast approximations of activation functions are used in deep learning[92].

Training algorithm. Once the architecture of the feed-forward ANN model is ready, now we need to train the model. For getting the training data, we divided the complete data in a 60:10:30 ratio randomly using the Mersenne Twister generator for training, validation, and testing, respectively. During the training phase, we use the training algorithm to minimise the output error by updating the weights and biases. We have used the Levenberg-Marquardt (LM) backpropagation algorithm for training the feed-forward ANN model[93]. It is based on Newton's method that was developed to optimise the sums of squares of the nonlinear functions. The Gauss-Newton approach itself has a limitation: the matrix might not be invertible. This problem can be resolved by modifying the Hessian matrix according to Eq. (14).

$$G = H + \mu I \tag{14}$$

where H represents the Hessian matrix, μ is a scalar that is the co-efficient of the steepest descent approach and Gauss-Newton method, and I is the identity matrix. To minimise the complexity involved in the computation of the Hessian matrix, it is approximated by the Jacobian matrix, which is computationally less expensive according to Eq. (15). The LM updates the values of weights and biases through a Newton-like iterative approach given by Eq. (16).

$$H = J^T J \tag{15}$$

$$w_{k+1} = w_k - \left[J^T J + \mu I\right]^{-1} J^T e \tag{16}$$

where J is the Jacobian matrix and e is the network error vector.

Results
Feature importance and association. We plot the relative importance score of each feature (Fig. 6a). A high value of the feature importance indicates more predictive power (i.e., more relevant feature). We observed that Longitude, VV, and VH have a high feature importance score. A high contribution from the backscatter features (i.e., VV and VH) is in accordance with previous studies[94]. High contribution from Longitude indicates a possible control of the geolocation feature due to the morphology of the Kosi Fan. This point has been elaborated in the "Discussion" section. Interestingly, the synthetic features (i.e., VH/VV and VH-VV) that were generated
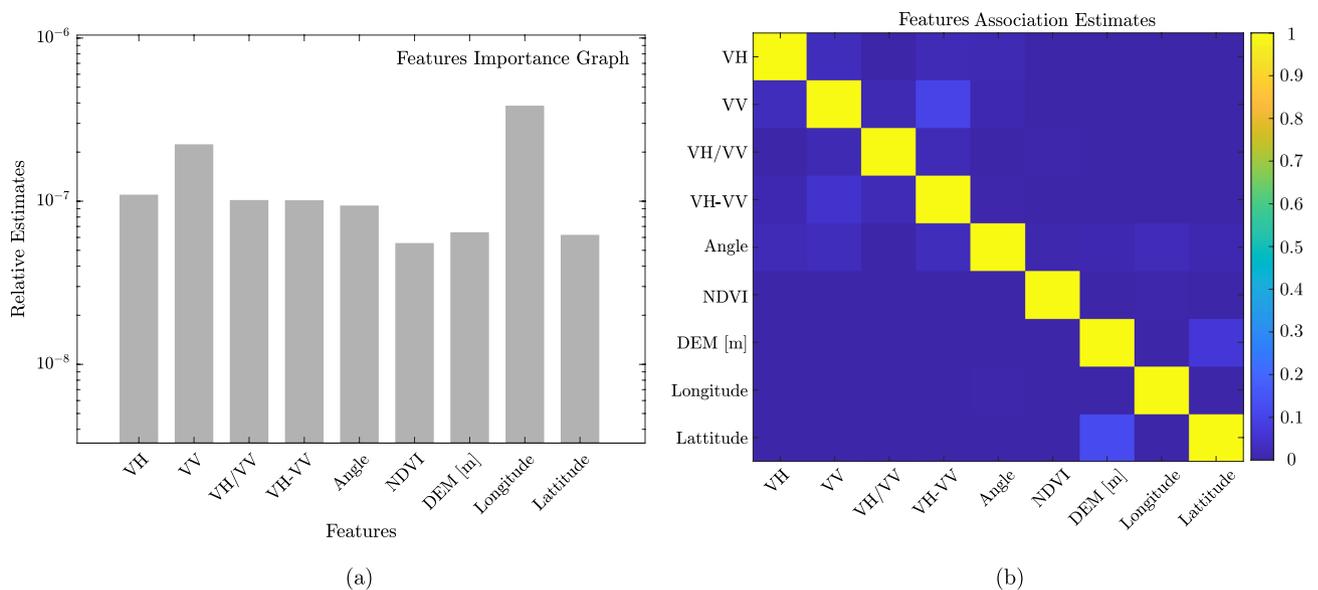
Figure 6. (**a**) The feature importance graph indicates the relative estimate of each feature obtained from the regression ensemble tree approach. The y-axis is in the log-scale. (**b**) Feature association graph indicating the correlation of each feature using heatmap.

through a linear data fusion have nearly the same importance score, which is relatively higher than the importance score of other input features such as incidence angle, NDVI, DEM, and Latitude. NDVI is relatively the least relevant feature with a minimum feature importance value.

Figure 6b is a heatmap of the feature association matrix. A high value in the feature association matrix indicates a high correlation among the features. We observed our features are not correlated. This indicates that the input features have appropriately trained the model without any instability and sensitivity.

Feature sensitivity. To analyse the impact of individual input features on predicting soil moisture, we have constructed the Partial Dependency Plot (PDP) and Individual Conditional Expectation (ICE) curves (Fig. 7). We do not observe a clear positive trend in VH. This could be due to the presence of some high-value VH observations (> 0.04) corresponding to low soil moisture (< 0.2 m^3/m^3). The presence of such limited oddity observations results in the dual behavior of VH, which is evident from the ICE curves. The majority of ICE curves corresponding to low-value VH (< 0.02) exhibit the same trend. However, we observed some ICE lines exhibit upward (after ≈ 0.025) while some (after ≈ 0.028) show downward trends. These trends get canceled out as the PDP takes the average of all the ICE lines, resulting a flat line after 0.03. The dual behavior in VH is generally observed, when soil moisture values are measured at the locations of high sub-pixel heterogeneity[95]. Such observations are inevitable while working with large and diverse in-situ measurements. The ICE curves of all other input features behave in a similar way without any significant deviation. This indicates that the PDP correctly illustrates the impact of all other features without concealing any local variations. We observed a strong positive impact of VV on soil moisture. VH/VV does not show any trend on soil moisture, whereas VH-VV has a slightly negative impact. The incidence angle and NDVI have undulated positive impact on soil moisture. An overall positive impact of DEM is observed. Longitude has a strong negative impact, whereas Latitude has a slight negative impact on soil moisture.

Performance of the machine learning model. To estimate the goodness of the training process, we assess the performance of our trained feed-forward ANN model over the training data. On training data, we discovered that the model performed reasonably well with R = 0.84, bias = 0 m^3/m^3, and RMSE = 0.04 m^3/m^3 (Fig. 8a). However, assessing the model performance solely on the training data results in a bias observation. We
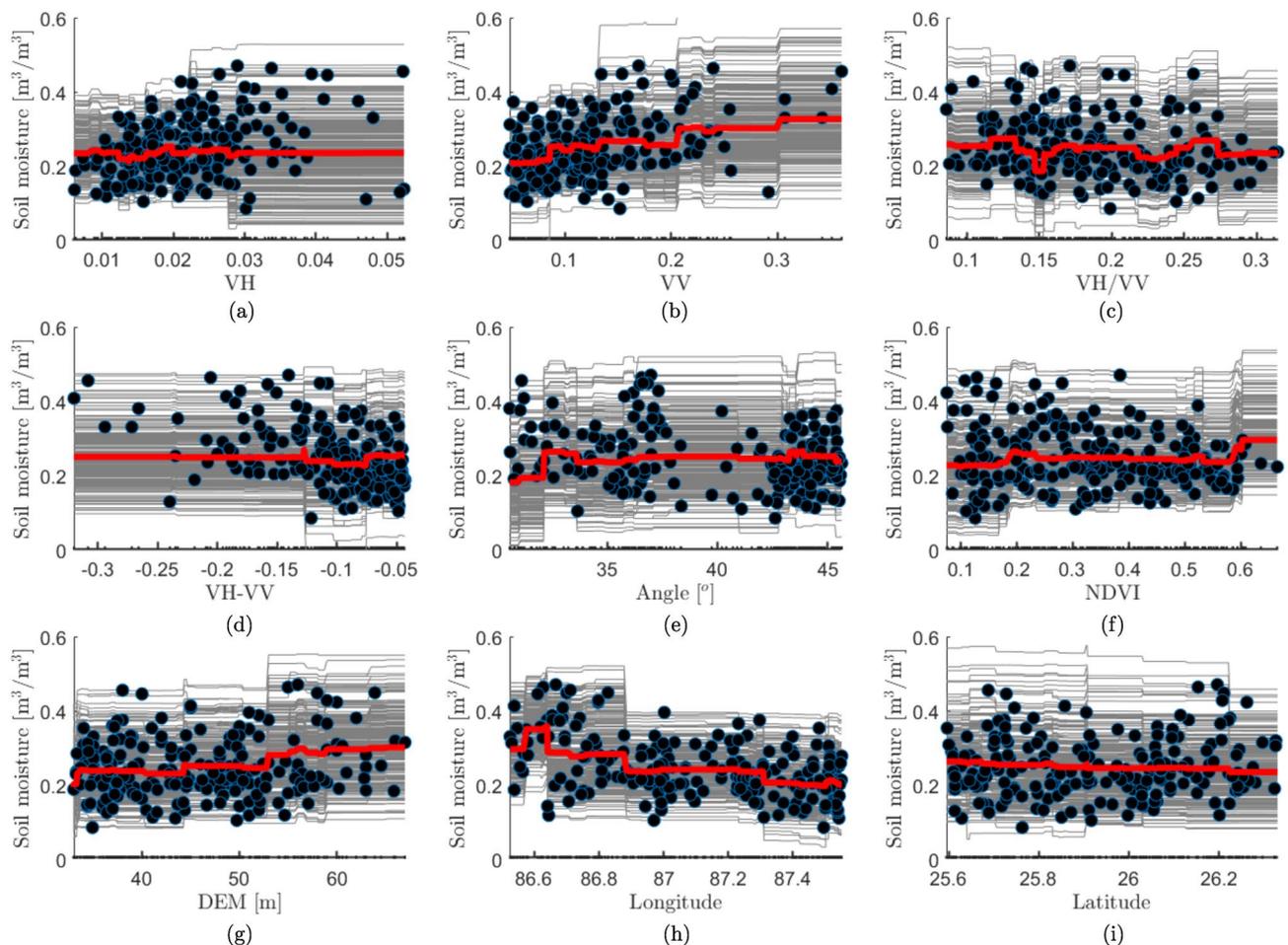
Figure 7. Feature sensitivity analysis using PDP (red line) and ICE curve (gray lines). The circles in black represent the observations.

need to evaluate the model performance by using unseen data (i.e., validation and testing data). We assess the model performance using the validation data while tuning the model parameters. We found a good agreement between the in-situ and predicted soil moisture for the validation process with R= 0.81, bias = 0.01 m^3/m^3, and RMSE = 0.05 m^3/m^3 (Fig. 8b). The presence of small positive bias indicates a slight overestimation of the trained model. We then fed the testing data into our model and evaluated the performance. We found the soil moisture measured in the field accorded well with the prediction with R = 0.80, bias = 0 m^3/m^3, RMSE = 0.04 m^3/m^3 (Fig. 8c). Finally, we report (Fig. 8d) the overall accuracy (R = 0.83, bias = 0 m^3/m^3, and RMSE = 0.04 m^3/m^3) of the trained model by evaluating its performance over the complete datasets (i.e., training+validation+testing).

Error and residual analysis. Figure 9 illustrates the error (i.e., predicted - in-situ) in the training, validation, and testing process. The height of the stacked bars represents the instances that occurred with the same error. We selected the bin size of twenty to represent the errors that range from − 0.1167 m^3/m^3 (leftmost bin) to +0.1091 m^3/m^3 (rightmost bin). The vertical orange line represents the zero error. The region left to this line represents the overestimation, and the region right to it is the underestimation. We have plotted a best fit Gaussian (curve in black) to the error histogram. Ideally, we expect the histogram to be normally distributed with zero mean. The distribution of our histogram is nearly normal, with the location of its peak at the zero error line indicating a good fit. Figure 10 shows the measured and predicted soil moisture at 95% confidence Interval (C.I). The predicted values of soil moisture accord well with the observed values. Further, to evaluate our model, we performed the residual analysis. We observed residuals are scattered randomly and do not show any pattern.
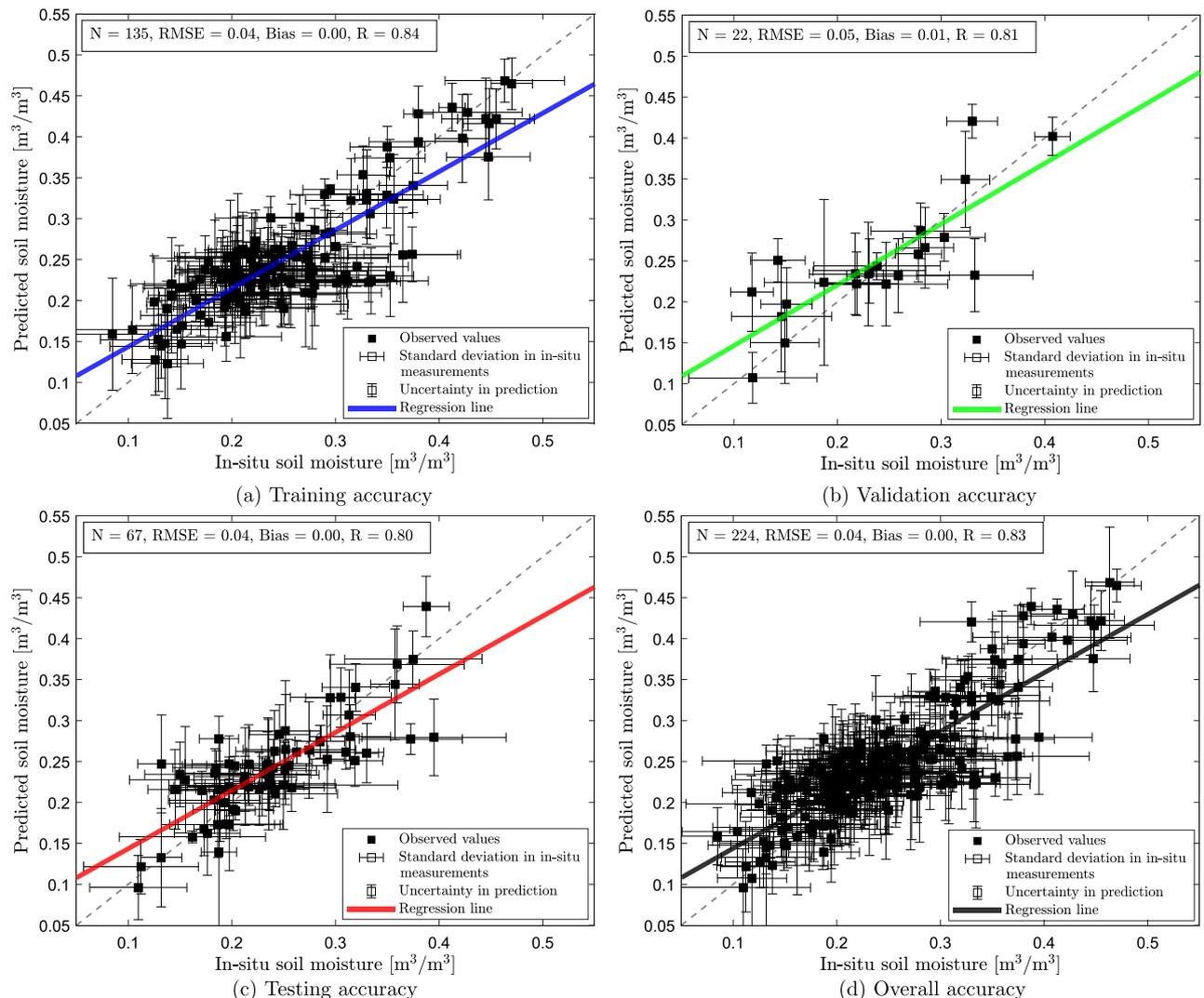
Figure 8. Model performance on (**a**) training, (**b**) validation, (**c**) testing, and (**d**) complete datasets. The dashed line in figure a-d represents y = x line. Horizontal error bar represents the standard deviation of the in-situ measurement, and vertical error bar represents the resulting uncertainty in the model prediction.
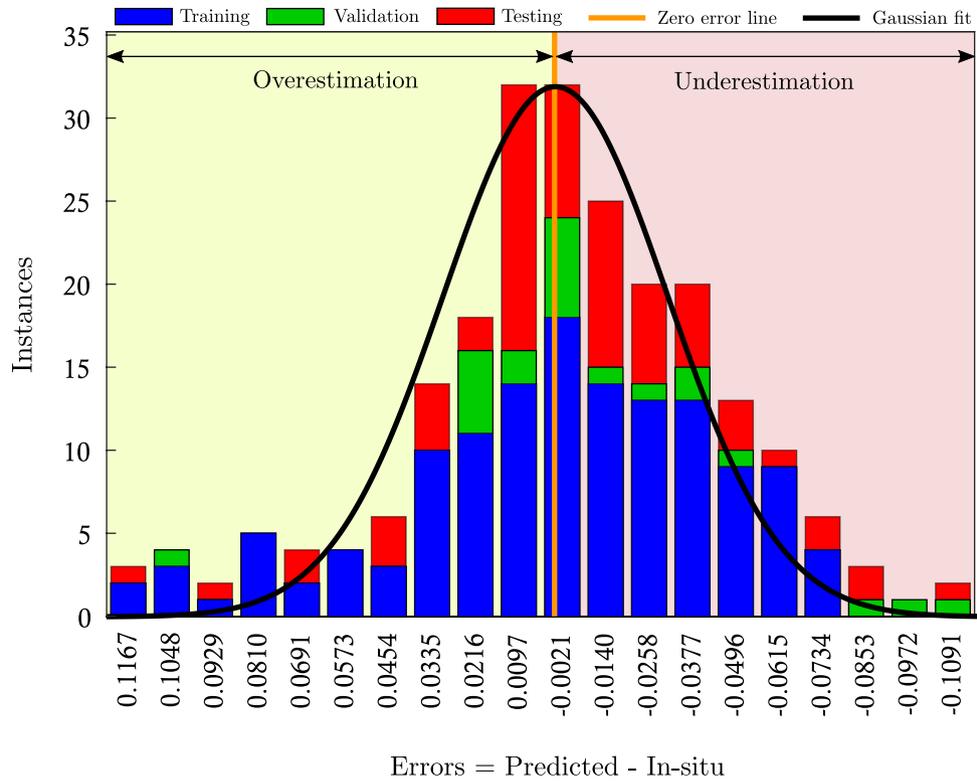
Figure 9. Error histogram (with 20 bin size) for training, validation, and testing phase. The area to the left and right of the zero error (orange) line represents overestimated and underestimated regions, respectively.

Finally, we performed the spatial distribution analysis. We formed thirty sets of training-validation-testing datasets by randomly dividing the in-situ observations (224) using Mersenne Twister random generator. We computed each set's training, validation, and testing accuracy using the proposed network and reported the $\mu \pm \sigma$. We observed an overall steady response for the training (R: 0.80 $\pm$0.05, RMSE: 0.05 $\pm$ 0.01 m^3/m^3, and bias: 0.00 $\pm$ 0.00 m^3/m^3), validation (R: 0.74 $\pm$ 0.08, RMSE: 0.05 $\pm$ 0.01 m^3/m^3, and bias: 0.00 $\pm$ 0.01 m^3/m^3), and testing accuracy (R: 0.72 $\pm$ 0.04, RMSE: 0.05 $\pm$ 0.01 m^3/m^3, and bias: 0.00 $\pm$ 0.01 m^3/m^3). This confirms the reliability and accuracy of the proposed network.
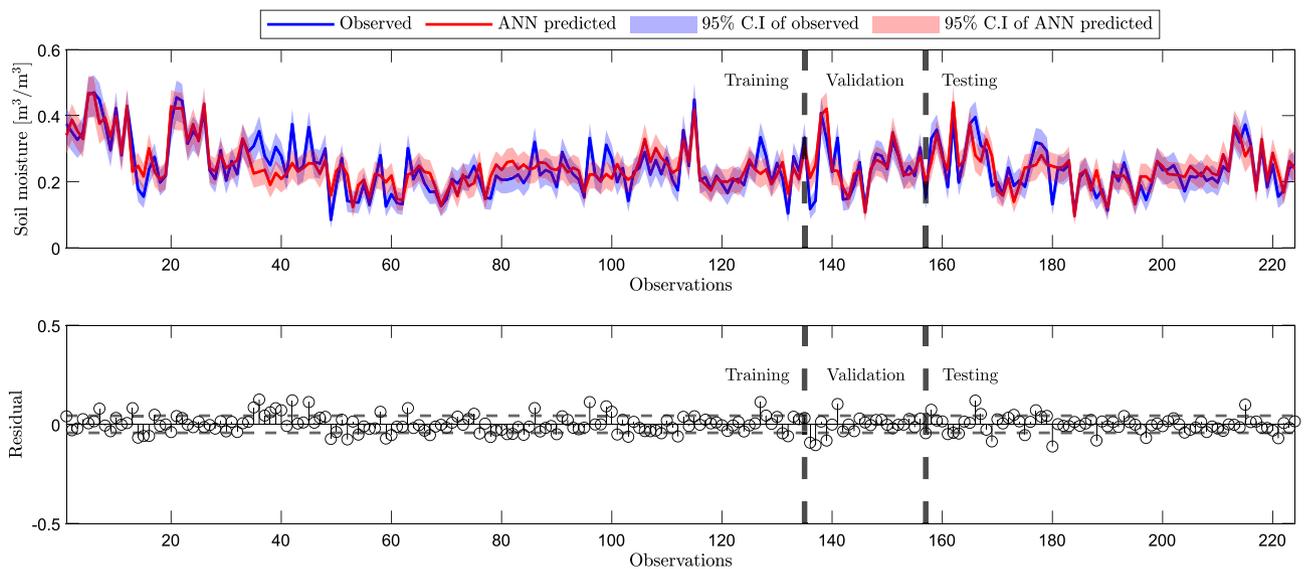
Figure 10. The top panel shows the line plot of observed and predicted soil moisture plotted for training, validation, and testing. The bottom panel shows the corresponding residual plot. The height of the vertical line with a circular cap represents the magnitude of the residual. The dashed line represents the overall RMSE value.

Discussion

Comparison with different scenarios of feed-forward ANN.   We have generated different scenarios based on two themes for robust analysis. For this, we vary the number of hidden layers and the respective neurons in each hidden layer. We have generated twelve scenarios by varying the number of hidden layers from one to four, with single, five, or ten neurons in each layer (Table 2). We evaluate the performance of these scenarios in the training, validation, and testing phase with three performance metrics (i.e., R, RMSE, and bias). We evaluated an additional performance metric, namely RMSE-observations standard deviation ratio (RSR), which is widely used for performance rating (very good, good, satisfactory, and unsatisfactory category) in the field of hydrology[96]. RSR (Eq. 17) consists of an error index with a normalisation factor, it can be used in diverse constituents. The value of RSR ranges from zero (indicating a perfect model with zero RMSE) to a large positive value (indicating a poor model with high RMSE).

$$RSR = \frac{Overall_{RMSE}}{STD_{in-situ}} \tag{17}$$

where $Overall_{RMSE}$ is the overall RMSE and $STD_{in-situ}$ is the standard deviation of the in-situ soil moisture measurements. We estimated the performance of the scenarios by different RSR categories; very good ($0 \leq RSR \leq 0.50$), good ($0.50 < RSR \leq 0.60$), satisfactory ($0.60 < RSR \leq 0.70$), and not satisfactory ($RSR > 0.70$). We observed that only the proposed scenario (i.e., 9-5-5-5-1) falls under the very good category. Out of twelve, two scenarios (i.e., 9-5-5-1 and 9-5-5-5-5-1) fall under the good category. Among these, we found that scenario 9-5-5-5-5-1 emerges as the best based on testing metrics. We have selected the best performing feed-forward ANN architecture (i.e., 9-5-5-5-1) to further generate ten different scenarios by varying the model inputs (Table 2). We observed that the performance of our architecture is optimal only when all nine features are considered.

Comparison with the benchmark algorithms and AutoML approach.   For a fair evaluation, we compared the performance of our fully connected feed-forward ANN with the other benchmark algorithms to

Scenarios		Training			Validation			Testing			RSR	Interpretation
		R	RMSE [m³/m³]	Bias [m³/m³]	R	RMSE [m³/m³]	Bias [m³/m³]	R	RMSE [m³/m³]	Bias [m³/m³]		
Layers/Neurons variation	9-1-1	0.68	0.06	0.003	0.75	0.06	− 0.017	0.62	0.06	− 0.003	0.74	Not satisfactory
	9-5-1	0.83	0.04	0.002	0.65	0.05	0.010	0.70	0.06	0.008	0.64	Satisfactory
	9-10-1	0.79	0.05	0.002	0.66	0.06	0.001	0.60	0.06	0.007	0.68	Satisfactory
	9-1-1-1	0.65	0.05	0.000	0.87	0.05	-0.017	0.67	0.06	-0.018	0.70	Satisfactory
	9-5-5-1	0.90	0.04	0.001	0.82	0.05	-0.017	0.59	0.05	0.009	0.58	Good
	9-10-10-1	0.78	0.05	0.007	0.73	0.06	0.010	0.57	0.06	0.013	0.70	Satisfactory
	9-1-1-1-1	0.71	0.05	0.000	0.85	0.06	0.001	0.52	0.06	0.009	0.74	Not satisfactory
	9-5-5-5-1	0.84	0.04	0.001	0.81	0.05	0.011	0.80	0.04	0.004	0.50	Very good
	9-10-10-10-1	0.74	0.05	0.000	0.85	0.04	0.000	0.77	0.05	0.019	0.63	Satisfactory
	9-1-1-1-1-1	0.68	0.06	− 0.002	0.61	0.05	0.011	0.65	0.06	− 0.008	0.75	Not satisfactory
	9-5-5-5-5-1	0.81	0.04	0.007	0.88	0.04	0.003	0.73	0.06	0.007	0.60	Good
	9-10-10-10-10-1	0.86	0.04	− 0.005	0.56	0.05	− 0.014	0.53	0.06	-0.018	0.67	Satisfactory
Input features variation	VH	0.39	0.07	-0.001	0.28	0.08	− 0.014	0.33	0.07	-0.003	0.93	Not satisfactory
	VV	0.29	0.08	0.008	0.45	0.07	0.004	0.49	0.07	0.014	0.93	Not satisfactory
	VH, VV	0.46	0.07	0.018	0.45	0.08	0.008	0.03	0.08	0.019	0.94	Not satisfactory
	VH/VV	0.16	0.08	0.006	0.27	0.08	0.000	0.27	0.07	0.011	0.98	Not satisfactory
	VH-VV	0.48	0.07	0.001	0.41	0.07	− 0.000	0.62	0.06	− 0.001	0.87	Not satisfactory
	VH, VV, VH/VV, VH-VV	0.58	0.06	0.002	0.36	0.08	− 0.016	0.18	0.07	− 0.001	0.89	Not satisfactory
	VH, VV, VH/VV, VH-VV, angle	0.68	0.06	-0.009	0.38	0.07	− 0.004	0.46	0.08	0.002	0.82	Not satisfactory
	VH, VV, VH/VV, VH-VV, angle, NDVI	0.68	0.06	− 0.002	0.39	0.06	0.014	0.50	0.06	0.009	0.78	Not satisfactory
	VH, VV, VH/VV, VH-VV, angle, NDVI, DEM	0.82	0.04	0.001	0.64	0.07	-0.027	0.56	0.07	0.007	0.70	Satisfactory
	All nine features	0.84	0.04	0.001	0.81	0.05	0.011	0.80	0.04	0.004	0.50	Very good

Table 2. Comparison of different scenarios by varying input feature combination and network architecture (hidden layers with neurons).

Performance metrics	Machine learning models										
	Feed forward ANN	GRNN	RBN	ERBN	GPR	SVR	RF	Boosting EL (LSBoost)	RNN	BDT	AutoML
R	0.80	0.24	0.18	0.12	0.28	0.13	0.19	0.16	0.24	0.19	0.37
RMSE [m^3/m^3]	0.040	0.060	0.061	0.062	0.060	0.062	0.061	0.061	0.059	0.048	0.048
Bias [m^3/m^3]	0.004	− 0.014	− 0.006	− 0.048	− 0.026	− 0.001	− 0.011	− 0.037	0.004	0.024	− 0.023

Table 3. Comparison with the benchmark algorithms (GRNN, RBN, Exact RBN, GPR, SVR, RF, Boosting EL, RNN, and BDT) and AutoML approach.

predict soil moisture by using the same data. We used the GRNN, RBN, ERBN, GPR, SVR, RF, Boosting EL, RNN, and BDT as the potential benchmark algorithms[97–106]. Other than these benchmark algorithms, we have also compared our result with the recently emerged Automated Machine Learning (AutoML) model[107]. We fed the same datasets into the AutoML platform of MATLAB® driven by *fitrauto* library. It automatically selects the machine learning model (i.e., linear regression, SVR, GPR, BDT, and EL) and optimise the corresponding tuning parameters through the Bayesian optimisation technique. During the optimisation process, it minimises the objective function (log(1 + CV$_{MSE}$); where CV$_{MSE}$ is the cross-validation MSE) iteratively. We found that the proposed ANN architecture outperforms all the benchmark algorithms with R = 0.80, bias = 0.004 m^3/m^3, and RMSE = 0.040 m^3/m^3 (Table 3).

To perform a robust and accurate comparison, we performed a statistical significance analysis to measure the performance of different ML models. To do so, we calculated the error in the predicted soil moisture (i.e., predicted - in-situ) for all the ML models. We applied statistical tests (i.e., Kolmogorov Smirnov and Shapiro-Wilk/Francia) to check the normality of the errors for each ML model. We found that the errors in each model are normally distributed. We then performed one-way ANOVA (ANalysis Of VAriance) to test a null hypothesis (i.e., h$_o$: mean of the error distribution of all ML models are equal). Based on the result of the ANOVA test, we rejected the null hypothesis. We noticed that two ML models (i.e., ERBN and Boosting EL) have significantly different mean values from the feed-forward ANN. In addition, we found that the RNN and feed-forward ANN have nearly the same mean (i.e., not significantly different), indicating similar performance. This is in accordance with the interpretation we drew from the analysis of performance metrics (i.e., bias).

We have evaluated the computational time complexity (using CPU 64-GB, @3.3 GHz, 10-cores) of the feed-forward ANN model and compared it with the benchmark algorithms (Fig. 11). The computational time complexity of a fully connected feed-forward ANN is $\mathcal{O}(\alpha n_{l1} + n_{l1}n_{l2} + \cdots)$, where α is the number of features and n_{li} is the number of neurons present at layer i[108]. We plotted the average computation time taken by each algorithm. We observed a clear trade-off between model performance and computational time complexity among these algorithms. The computational time complexity of the feed-forward ANN is slightly higher than the other algorithms. This is probably due to the large number of computations involved during the optimisation of the hyperparameters (i.e., weights and biases) by using the LM backpropagation algorithm.

Sensitivity analysis. Sensitivity analysis is important to assess the consistency of a data-driven model. We evaluated the response of our feed-forward ANN model concerning the uncertainty in the input features. In doing so, we introduce ±5%, and ±10% uncertainty in all the input features at a time by keeping other features unchanged and evaluating how these uncertainties from individual features contribute to the total uncertainty in the response variable (i.e., soil moisture). We observed for ±5%, and ±10% uncertainty in the input features, the uncertainty in the model-derived soil moisture ranges between ≈ − 4% and ≈ + 8% (Fig. 12). The model is more resistant to the presence of uncertainty in VH, VH-VV, incidence angle, and Latitude. In contrast, it is less resistant to the presence of uncertainty in NDVI and Longitude.

Soil moisture on the Kosi Fan. Figures 13a and b are the surface soil moisture maps generated from the ANN for two different time frames (i.e., 17 December 2019 and 06 March 2022). The invalid regions (dense built-up and water bodies) have been masked.

On the Kosi Fan, soil moisture appears relatively high at the western margin. It is important to note that the Kosi River flows at the western margin. To understand the spatial variation of soil moisture, we performed a topographic analysis. A transverse (TT') transect drown on a DEM exhibits a convex-upward profile of the Kosi Fan (Fig. 14). The elevation is maximum at the fan axis and decreases towards the western and eastern margins of the fan. This inherent topography of fan controls the drainage organisation. The drainage networks diverge from the fan axis towards the fan margins. Also, the groundwater table appears to follow surface topography. The water table is at a shallow depth in close proximity to the Kosi River (Fig. 14c). The topography, drainage orientation, and shallow water table at the western margin of the fan make this region prone to high soil moisture content.

Further, to assess the spatio-temporal variation of soil moisture content, we take the difference of soil moisture maps of two different dates (Fig. 13c). We observed high variability in the moisture content. This suggests a strong seasonal variability of soil moisture content on the fan surface. Processing long-time series microwave images would help to quantify the inter-annual variation and also to understand the impact of climate change and human perturbation on the spatio-temporal dynamics of soil moisture. Such analysis would be very useful to predict soil moisture conditions in the near future, which would help to plan agriculture and food security in the region.
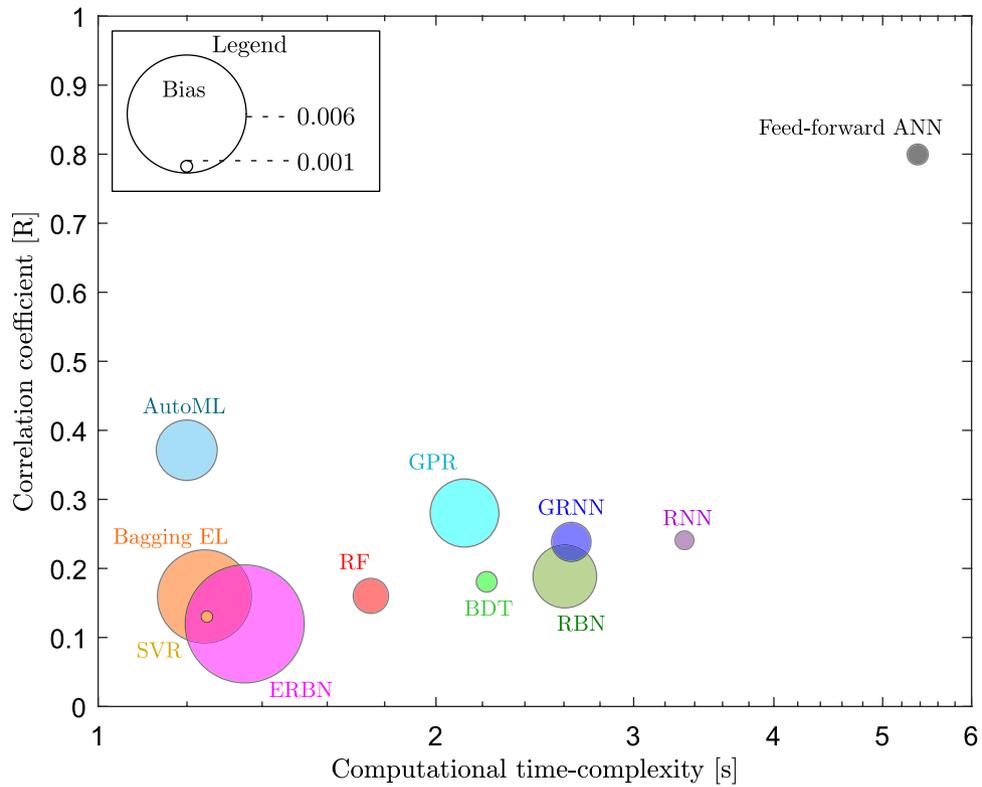
Figure 11. Comparison of the computational time complexity and performance of the benchmark algorithms (each represented in a different colour). The radius of the circle represents the magnitude of the bias for each model.
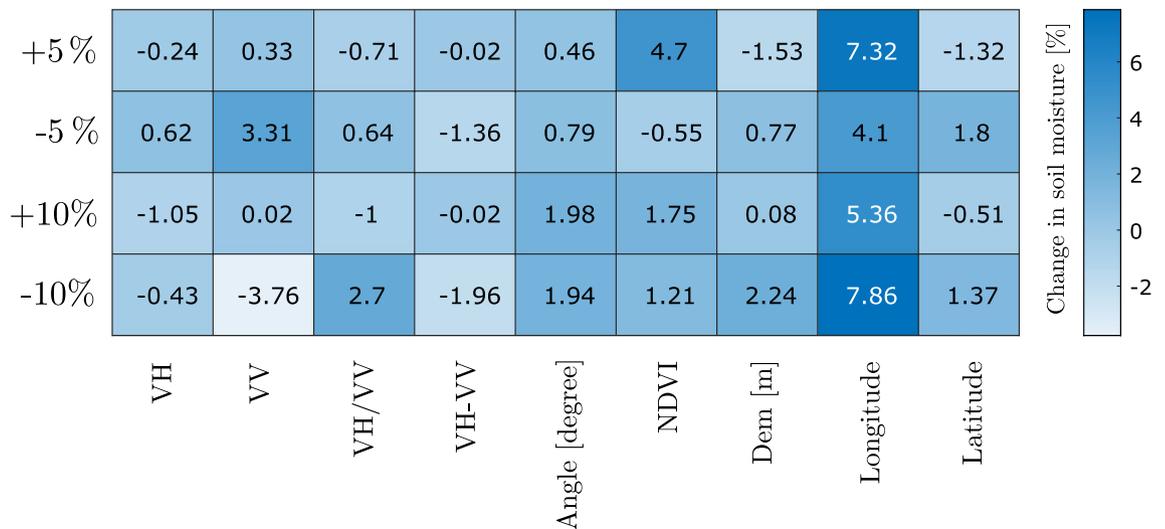
Figure 12. Sensitivity of feed-forward ANN architecture (9-5-5-5-1) by considering uncertainties (∓ 5% and ∓ 10%) in input features.

Conclusion

We applied a fully connected feed-forward (i.e., 9-5-5-5-1) ANN algorithm and data fusion to estimate surface soil moisture on the Kosi alluvial fan using the multi-sensor remote sensing images. From the input features, the Longitude, VV, and VH have emerged as the most relevant features for mapping surface soil moisture. Among these, Longitude and VV exhibit negative and positive impacts on soil moisture, respectively. We do not observe a clear positive trend for VH. Since the in-situ measurement of larger sample size usually contains few oddity samples, primarily from a location of large sub-pixel heterogeneity and dense vegetation. These samples usually
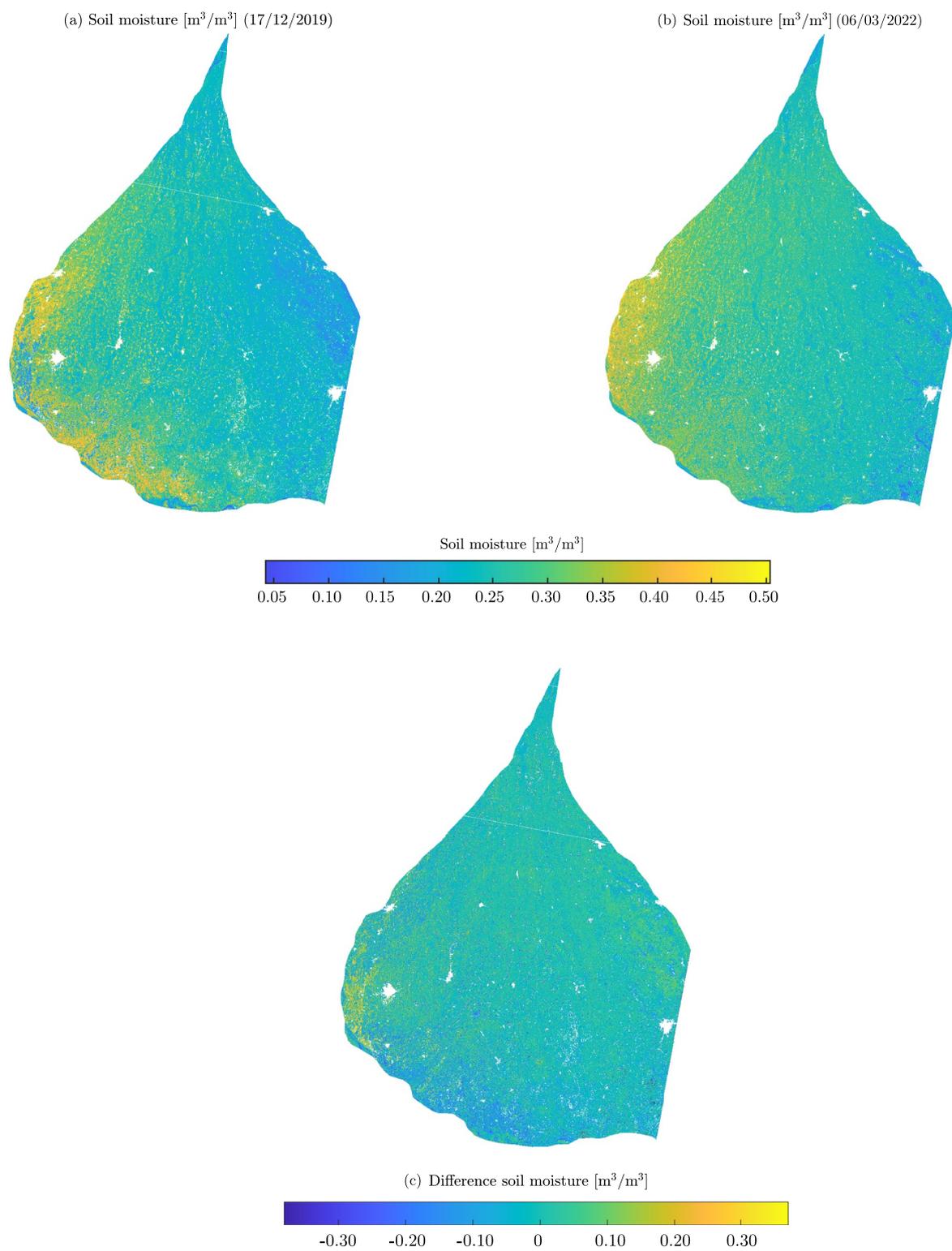
Figure 13. (**a**) High spatial resolution (i.e., 60 m) surface soil moisture map for 11th December 2019 and (**b**) 06th March 2022 (**c**) Corresponding difference soil moisture map. The transparent pixels represent the invalid regions.
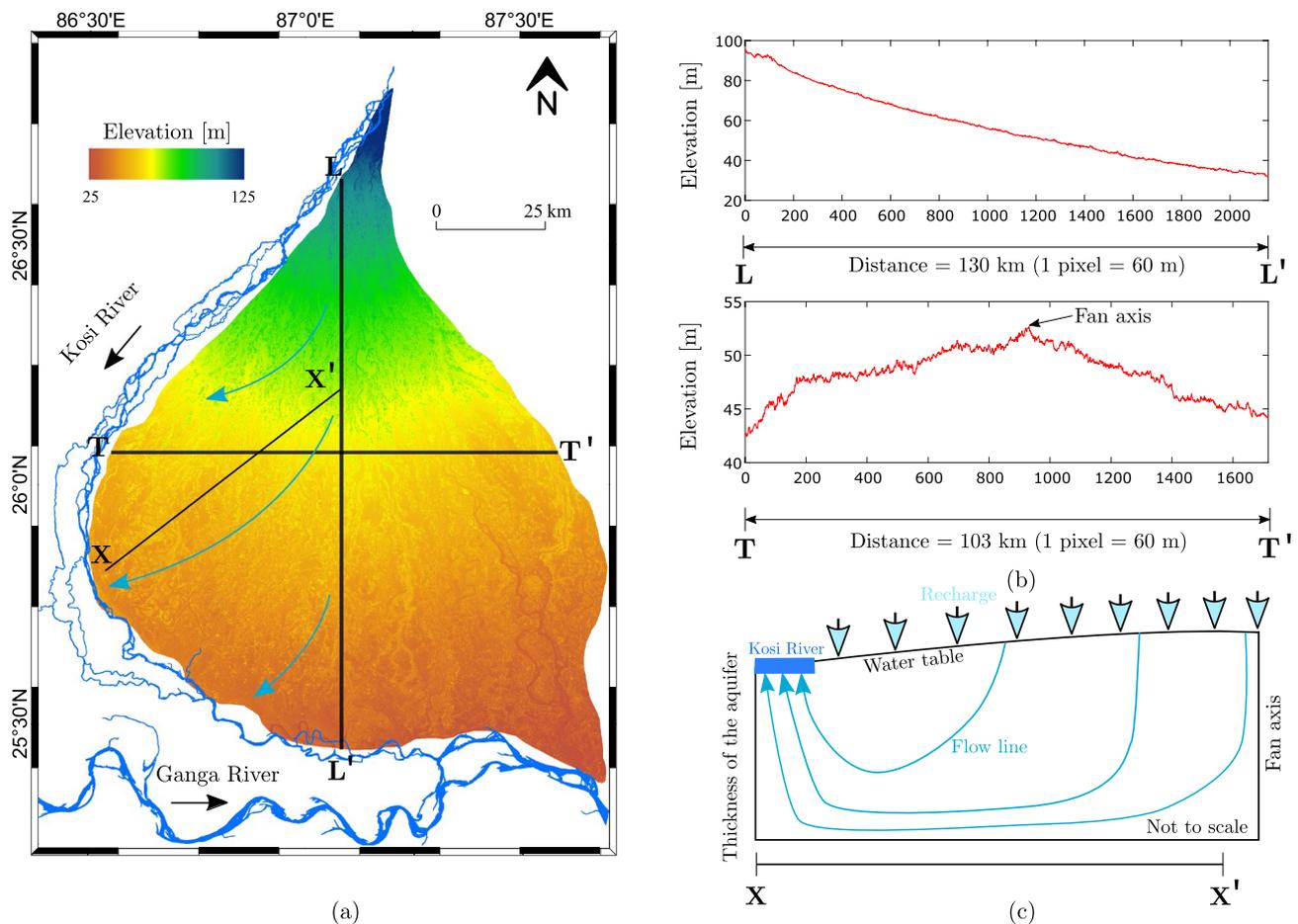
Figure 14. (**a**) Digital elevation model of the Kosi Fan, (**b**) elevation profiles on the fan along the longitudinal (LL′) and transverse (TT′) transects, and (**c**) schematic of the groundwater flow lines from the fan axis towards the western margin of the fan (modified after Khan et al.[109]).

return high VH values for low soil moisture resulting in a dual behaviour of VH[95]. We noticed DEM has a clear positive impact on soil moisture, whereas Latitude has a negative impact on soil moisture. The incidence angle and NDVI have a fluctuating positive impact, and VH-VV has a fluctuating negative impact on soil moisture. VH/VV did not show any clear trend. The spatial pattern of the surface soil moisture over Kosi Fan indicates a possible control of surface topology and fan morphology.

ANN with three hidden layers having five neurons (i.e., 9-5-5-5-1) each has a relatively high predictability of surface soil moisture than the other benchmark algorithms. However, there is a trade-off between the performance and computational time complexity. The fully connected feed-forward ANN has the highest time complexity with the best performance. This model is relatively more sensitive towards the presence of small uncertainly in the graphical indicator and geolocation features (i.e., NDVI and Longitude) than other input features.

This comprehensive framework allows us to generate a surface soil moisture map from dual polarised backscatter images from Sentinel-1, red and near-infrared surface reflectance from Sentinel-2, and DEM from SRTM satellite images. The outcome of this study could be used as input data to study waterlogging, flood inundation, agronomy, drainage congestion, drought prediction, and other hydrological applications.

Code availability
The computer algorithms originated during the current study can be made available from the corresponding author on a reasonable request.

Data availibility
The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

References
1. McColl, K. A. *et al.* The global distribution and dynamics of surface soil moisture. *Nat. Geosci.* **10**, 100–104 (2017).
2. Babaeian, E. *et al.* Ground, proximal, and satellite remote sensing of soil moisture. *Rev. Geophys.* **57**, 530–616 (2019).

3. Njoku, E. G. & Entekhabi, D. Passive microwave remote sensing of soil moisture. *J. Hydrol.* **184**, 101–129 (1996).
4. Ahmad, S., Kalra, A. & Stephen, H. Estimating soil moisture using remote sensing data: A machine learning approach. *Adv. Water Resour.* **33**, 69–80 (2010).
5. Mohanty, B. P., Cosh, M. H., Lakshmi, V. & Montzka, C. Soil moisture remote sensing: State-of-the-science. *Vadose Zone J.* **16**, 1–9 (2017).
6. Balenzano, A. *et al.* Sentinel-1 soil moisture at 1 km resolution: A validation study. *Remote Sens. Environ.* **263**, 112554 (2021).
7. Kerr, Y. H. *et al.* Soil moisture retrieval from space: The soil moisture and ocean salinity (smos) mission. *IEEE Trans. Geosci. Remote Sens.* **39**, 1729–1735 (2001).
8. Entekhabi, D. *et al.* The soil moisture active passive (smap) mission. *Proc. IEEE* **98**, 704–716 (2010).
9. Emery, W. & Camps, A. *Introduction to Satellite Remote Sensing: Atmosphere, Ocean, Land and Cryosphere Applications* (Elsevier, 2017).
10. Baldwin, D., Manfreda, S., Lin, H. & Smithwick, E. A. Estimating root zone soil moisture across the eastern united states with passive microwave satellite data and a simple hydrologic model. *Remote Sens.* **11**, 2013 (2019).
11. Sungmin, O. & Orth, R. Global soil moisture data derived through machine learning trained with in-situ measurements. *Sci. Data* **8**, 1–14 (2021).
12. Bauer-Marschallinger, B. *et al.* Soil moisture from fusion of scatterometer and sar: Closing the scale gap with temporal filtering. *Remote Sens.* **10**, 1030 (2018).
13. Zribi, M. *et al.* Analysis of l-band sar data for soil moisture estimations over agricultural areas in the tropics. *Remote Sens.* **11**, 1122 (2019).
14. Singh, A., Gaurav, K., Meena, G. K. & Kumar, S. Estimation of soil moisture applying modified dubois model to sentinel-1; a regional study from central india. *Remote Sens.* **12**, 2266 (2020).
15. Altese, E., Bolognani, O., Mancini, M. & Troch, P. A. Retrieving soil moisture over bare soil from ers 1 synthetic aperture radar data: Sensitivity analysis based on a theoretical surface scattering model and field data. *Water Resour. Res.* **32**, 653–661 (1996).
16. Ulaby, F. Microwave remote sensing; fundamentals and radiometory. *Microwave Rem. Sens. Active Passive* **1**, 256 (1981).
17. Barrett, B. W., Dwyer, E. & Whelan, P. Soil moisture retrieval from active spaceborne microwave observations: An evaluation of current techniques. *Remote Sens.* **1**, 210–242 (2009).
18. Oh, Y. *et al.* An empirical model and an inversion technique for radar scattering from bare soil surfaces. *IEEE Trans. Geosci. Remote Sens.* **30**, 370–381 (1992).
19. Oh, Y., Sarabandi, K. & Ulaby, F. T. Semi-empirical model of the ensemble-averaged differential mueller matrix for microwave backscattering from bare soil surfaces. *IEEE Trans. Geosci. Remote Sens.* **40**, 1348–1355 (2002).
20. Oh, Y. Quantitative retrieval of soil moisture content and surface roughness from multipolarized radar observations of bare soil surfaces. *IEEE Trans. Geosci. Remote Sens.* **42**, 596–601 (2004).
21. Dubois, P. C., Van Zyl, J. & Engman, T. Measuring soil moisture with imaging radars. *IEEE Trans. Geosci. Remote Sens.* **33**, 915–926 (1995).
22. Fung, A. K., Li, Z. & Chen, K.-S. Backscattering from a randomly rough dielectric surface. *IEEE Trans. Geosci. Remote Sens.* **30**, 356–369 (1992).
23. Attema, E. & Ulaby, F. T. Vegetation modeled as a water cloud. *Radio Sci.* **13**, 357–364 (1978).
24. Topp, G. C., Davis, J. & Annan, A. P. Electromagnetic determination of soil water content: Measurements in coaxial transmission lines. *Water Resour. Res.* **16**, 574–582 (1980).
25. Walker, J. P. & Troch, P. A. Profile soil moisture estimation using the modified iem. In *IGARSS'97. 1997 IEEE International Geoscience and Remote Sensing Symposium Proceedings. Remote Sensing-A Scientific Vision for Sustainable Development*, vol. 3 1263–1265 (IEEE, 1997).
26. Bindlish, R. & Barros, A. P. Multifrequency soil moisture inversion from sar measurements with the use of iem. *Remote Sens. Environ.* **71**, 67–88 (2000).
27. Zribi, M. & Dechambre, M. A new empirical model to retrieve soil moisture and roughness from c-band radar data. *Remote Sens. Environ.* **84**, 42–52 (2003).
28. Baghdadi, N. & Zribi, M. Evaluation of radar backscatter models iem, oh and dubois using experimental observations. *Int. J. Remote Sens.* **27**, 3831–3852 (2006).
29. Baghdadi, N., Saba, E., Aubert, M., Zribi, M. & Baup, F. Evaluation of radar backscattering models iem, oh, and dubois for sar data in x-band over bare soils. *IEEE Geosci. Remote Sens. Lett.* **8**, 1160–1164 (2011).
30. Panciera, R., Tanase, M. A., Lowell, K. & Walker, J. P. Evaluation of iem, dubois, and oh radar backscatter models using airborne l-band sar. *IEEE Trans. Geosci. Remote Sens.* **52**, 4966–4979 (2013).
31. Choker, M. *et al.* Evaluation of the oh, dubois and iem backscatter models using a large dataset of sar data and experimental soil measurements. *Water* **9**, 38 (2017).
32. Mirsoleimani, H. R., Sahebi, M. R., Baghdadi, N. & El Hajj, M. Bare soil surface moisture retrieval from sentinel-1 sar data based on the calibrated iem and dubois models using neural networks. *Sensors* **19**, 3209 (2019).
33. Weiß, T., Ramsauer, T., Löw, A. & Marzahn, P. Evaluation of different radiative transfer models for microwave backscatter estimation of wheat fields. *Remote Sens.* **12**, 3037 (2020).
34. Rao, S. S. *et al.* Modified dubois model for estimating soil moisture with dual polarized sar data. *J. Indian Soc. Remote Sens.* **41**, 865–872 (2013).
35. Shen, X., Mao, K., Qin, Q., Hong, Y. & Zhang, G. Bare surface soil moisture estimation using double-angle and dual-polarization l-band radar data. *IEEE Trans. Geosci. Remote Sens.* **51**, 3931–3942 (2013).
36. Kweon, S.-K. & Oh, Y. Estimation of soil moisture and surface roughness from single-polarized radar data for bare soil surface and comparison with dual-and quad-polarization cases. *IEEE Trans. Geosci. Remote Sens.* **52**, 4056–4064 (2013).
37. Thanabalan, P. & Vidhya, R. A synergistic approach for soil moisture estimation using modified dubois model with dual-polarized sar and optical satellite data. In *Land Surface and Cryosphere Remote Sensing III*, vol. 9877, 98770Y (International Society for Optics and Photonics, 2016).
38. Dave, R., Kumar, G., Pandey, D., Khan, A. & Bhattacharya, B. Evaluation of modified dubois model for estimating surface soil moisture using dual polarization risat-1 c-band sar data. *Geocarto Int.* **36**, 1459–1469 (2021).
39. Singh, A., Gaurav, K., Rai, A. K. & Beg, Z. Machine learning to estimate surface roughness from satellite images. *Remote Sens.* **13**, 3794 (2021).
40. Wagner, W. *et al.* Operational readiness of microwave remote sensing of soil moisture for hydrologic applications. *Hydrol. Res.* **38**, 1–20 (2007).
41. Hsieh, C.-Y., Fung, A. K., Nesti, G., Sieber, A. J. & Coppo, P. A further study of the iem surface scattering model. *IEEE Trans. Geosci. Remote Sens.* **35**, 901–909 (1997).
42. Mancini, M., Hoeben, R. & Troch, P. A. Multifrequency radar observations of bare surface soil moisture content: A laboratory experiment. *Water Resour. Res.* **35**, 1827–1838 (1999).
43. Lee, C. S., Sohn, E., Park, J. D. & Jang, J.-D. Estimation of soil moisture using deep learning based on satellite data: A case study of south korea. *GIScience Remote Sens.* **56**, 43–67 (2019).
44. Cai, Y., Zheng, W., Zhang, X., Zhangzhong, L. & Xue, X. Research on soil moisture prediction model based on deep learning. *PLoS ONE* **14**, e0214508 (2019).

45. Chaudhary, S. K. *et al.* Machine learning algorithms for soil moisture estimation using sentinel-1: Model development and implementation. *Adv. Space Res.* **69**, 1799–1812 (2022).
46. Xu, H. *et al.* Quality improvement of satellite soil moisture products by fusing with in-situ measurements and gnss-r estimates in the western continental us. *Remote Sens.* **10**, 1351 (2018).
47. Yuan, Q., Xu, H., Li, T., Shen, H. & Zhang, L. Estimating surface soil moisture from satellite observations using a generalized regression neural network trained on sparse ground-based measurements in the continental us. *J. Hydrol.* **580**, 124351 (2020).
48. Del Frate, F., Ferrazzoli, P. & Schiavon, G. Retrieving soil moisture and agricultural variables by microwave radiometry using neural networks. *Remote Sens. Environ.* **84**, 174–183 (2003).
49. Elshorbagy, A. & Parasuraman, K. On the relevance of using artificial neural networks for estimating soil moisture content. *J. Hydrol.* **362**, 1–18 (2008).
50. Santi, E., Paloscia, S., Pettinato, S. & Fontanelli, G. Application of artificial neural networks for the soil moisture retrieval from active and passive microwave spaceborne sensors. *Int. J. Appl. Earth Obs. Geoinf.* **48**, 61–73 (2016).
51. Santi, E. *et al.* Soil moisture and forest biomass retrieval on a global scale by using cygnss data and artificial neural networks. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium* 5905–5908 (IEEE, 2020).
52. Santi, E., Dabboor, M., Pettinato, S. & Paloscia, S. Combining machine learning and compact polarimetry for estimating soil moisture from c-band sar data. *Remote Sens.* **11**, 2451 (2019).
53. Adab, H., Morbidelli, R., Saltalippi, C., Moradian, M. & Ghalhari, G. A. F. Machine learning to estimate surface soil moisture from remote sensing data. *Water* **12**, 3223 (2020).
54. Datta, S., Das, P., Dutta, D. & Giri, R. K. Estimation of surface moisture content using sentinel-1 c-band sar data through machine learning models. *J. Indian Soc. Remote Sens.* **49**, 887–896 (2021).
55. Greifeneder, F., Notarnicola, C. & Wagner, W. A machine learning-based approach for surface soil moisture estimations with google earth engine. *Remote Sens.* **13**, 2099 (2021).
56. Gole, C. V. & Chitale, S. V. Inland delta building activity of kosi river. *J. Hydraul. Div.* **92**, 111–126 (1966).
57. Wells, N. A. & Dorr, J. A. Jr. Shifting of the kosi river, northern india. *Geology* **15**, 204–207 (1987).
58. Sinha, R. The great avulsion of kosi on 18 august 2008. *Curr. Sci.* **2009**, 429–433 (2009).
59. Gaurav, K. *et al.* Morphology of the kosi megafan channels. *Earth Surf. Dyn.* **3**, 321–331 (2015).
60. Agarwal, R. & Bhoj, R. Evolution of kosi river fan, India: Structural implications and geomorphic significance. *Int. J. Remote Sens.* **13**, 1891–1901 (1992).
61. Gaurav, K., Tandon, S., Devauchelle, O., Sinha, R. & Métivier, F. A single width-discharge regime relationship for individual threads of braided and meandering rivers from the himalayan foreland. *Geomorphology* **295**, 126–133 (2017).
62. IMD. (2011; accessed 25 March 2021); *Climate of Bihar*.
63. Kumar, A., Sarthi, P. P., Kumari, A. & Sinha, A. K. Observed characteristics of rainfall indices and outgoing longwave radiation over the gangetic plain of india. *Pure Appl. Geophys.* **178**, 619–631 (2021).
64. NRSC. (2017; accessed 25 March 2021)); *District and category wise distribution of Land Use and Land Cover in Bihar (2015–16)*.
65. DeVries, B. *et al.* Rapid and robust monitoring of flood events using sentinel-1 and landsat data on the google earth engine. *Remote Sens. Environ.* **240**, 111664 (2020).
66. Singh, A., Meena, G. K., Kumar, S. & Gaurav, K. Analysis of the effect of incidence angle and moisture content on the penetration depth of l-and s-band sar signals into the ground surface. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* 4 (2018).
67. Singh, A., Meena, G. K., Kumar, S. & Gaurav, K. Evaluation of the penetration depth of l-and s-band (nisar mission) microwave sar signals into ground. In *2019 URSI Asia-Pacific Radio Science Conference (AP-RASC)* 1 (IEEE, 2019).
68. Singh, A., Niranjannaik, M., Kumar, S. & Gaurav, K. Comparison of different dielectric models to estimate penetration depth of l-and s-band sar signals into the ground surface. *Geographies* **2**, 734–742 (2022).
69. Martins, V. S. *et al.* Assessment of atmospheric correction methods for sentinel-2 msi images applied to amazon floodplain lakes. *Remote Sens.* **9**, 322 (2017).
70. Li, J. & Roy, D. P. A global analysis of sentinel-2a, sentinel-2b and landsat-8 data revisit intervals and implications for terrestrial monitoring. *Remote Sens.* **9**, 902 (2017).
71. Nasa, J. Nasa shuttle radar topography mission global 1 arc second. *Nasa Lp Daac* **15**, 256 (2013).
72. Ryan, J. *et al.* How robust are in situ observations for validating satellite-derived albedo over the dark zone of the greenland ice sheet?. *Geophys. Res. Lett.* **44**, 6218–6225 (2017).
73. Thakur, K. K. *et al.* Comparison of remotely-sensed sea surface temperature and salinity products with in situ measurements from british columbia, canada. *Front. Mar. Sci.* **5**, 121 (2018).
74. Jensen, J. R. *et al.Introductory Digital Image Processing: A Remote Sensing Perspective*. Ed. 2 (Prentice-Hall Inc., 1996).
75. Autret, M., Bernard, R. & Vidal-Madjar, D. Theoretical study of the sensitivity of the microwave backscattering coefficient to the soil surface parameters. *Remote Sens.* **10**, 171–179 (1989).
76. Srivastava, H. S., Patel, P., Sharma, Y. & Navalgund, R. R. Large-area soil moisture estimation using multi-incidence-angle radarsat-1 sar data. *IEEE Trans. Geosci. Remote Sens.* **47**, 2528–2535 (2009).
77. Lin, H., Kogelmann, W., Walker, C. & Bruns, M. Soil moisture patterns in a forested catchment: A hydropedological perspective. *Geoderma* **131**, 345–368 (2006).
78. Tenenbaum, D., Band, L., Kenworthy, S. & Tague, C. Analysis of soil moisture patterns in forested and suburban catchments in baltimore, maryland, using high-resolution photogrammetric and lidar digital elevation datasets. *Hydrol. Process. Int. J.* **20**, 219–240 (2006).
79. Murphy, P., Ogilvie, J. & Arp, P. Topographic modelling of soil moisture conditions: A comparison and verification of two models. *Eur. J. Soil Sci.* **60**, 94–109 (2009).
80. Fathololoumi, S., Vaezi, A. R., Alavipanah, S. K., Ghorbani, A. & Biswas, A. Comparison of spectral and spatial-based approaches for mapping the local variation of soil moisture in a semi-arid mountainous area. *Sci. Total Environ.* **724**, 138319 (2020).
81. Martin, R., Aler, R., Valls, J. M. & Galván, I. M. Machine learning techniques for daily solar energy prediction and interpolation using numerical weather models. *Concurr. Comput. Pract. Exp.* **28**, 1261–1274 (2016).
82. Ahn, S., Ryu, D.-W. & Lee, S. A machine learning-based approach for spatial estimation using the spatial features of coordinate information. *ISPRS Int. J. Geo Inf.* **9**, 587 (2020).
83. Ittner, A. & Schlosser, M. Discovery of relevant new features by generating non-linear decision trees. *KDD* **1996**, 108–113 (1996).
84. Greifeneder, F. *et al.* The added value of the vh/vv polarization-ratio for global soil moisture estimations from scatterometer data. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **11**, 3668–3679 (2018).
85. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, 1189–1232 (2001).
86. Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.* **24**, 44–65 (2015).
87. Apicella, A., Donnarumma, F., Isgrò, F. & Prevete, R. A survey on modern trainable activation functions. *Neural Netw.* **138**, 14–32 (2021).
88. Hayou, S., Doucet, A. & Rousseau, J. On the impact of the activation function on deep neural networks training. In *International Conference on Machine Learning* 2672–2680 (PMLR, 2019).

89. Manni, A., Saviano, G. & Bonelli, M. G. Optimization of the anns predictive capability using the taguchi approach: A case study. *Mathematics* **9**, 766 (2021).
90. Roy, S. K., Manna, S., Dubey, S. R. & Chaudhuri, B. B. Lisht: Non-parametric linearly scaled hyperbolic tangent activation function for neural networks. arXiv:1901.05894 (2019).
91. Tan, T. G., Teo, J. & Anthony, P. A comparative investigation of non-linear activation functions in neural controllers for search-based game ai engineering. *Artif. Intell. Rev.* **41**, 1–25 (2014).
92. Cococcioni, M., Rossi, F., Ruffaldi, E. & Saponara, S. Fast approximations of activation functions in deep neural networks when using posit arithmetic. *Sensors* **20**, 1515 (2020).
93. Hagan, M. T. & Menhaj, M. B. Training feedforward networks with the marquardt algorithm. *IEEE Trans. Neural Netw.* **5**, 989–993 (1994).
94. Bousbih, S. *et al.* Potential of sentinel-1 radar data for the assessment of soil and cereal cover parameters. *Sensors* **17**, 2617 (2017).
95. Dabrowska-Zielinska, K. *et al.* Soil moisture in the biebrza wetlands retrieved from sentinel-1 imagery. *Remote Sens.* **10**, 1979 (2018).
96. Moriasi, D. N. *et al.* Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* **50**, 885–900 (2007).
97. Zhang, D., Zhang, N., Ye, N., Fang, J. & Han, X. Hybrid learning algorithm of radial basis function networks for reliability analysis. *IEEE Trans. Reliab.* **70**, 887–900 (2020).
98. Elias, I. *et al.* Genetic algorithm with radial basis mapping network for the electricity consumption modeling. *Appl. Sci.* **10**, 4239 (2020).
99. Lin, H. *et al.* Radial basis function artificial neural network able to accurately predict disinfection by-product levels in tap water: Taking haloacetic acids as a case study. *Chemosphere* **248**, 125999 (2020).
100. Mansor, M. *et al.* Systematic boolean satisfiability programming in radial basis function neural network. *Processes* **8**, 214 (2020).
101. Singh, A., Kotiyal, V., Sharma, S., Nagar, J. & Lee, C.-C. A machine learning approach to predict the average localization error with applications to wireless sensor networks. *IEEE Access* **8**, 208253–208263 (2020).
102. Singh, A., Nagar, J., Sharma, S. & Kotiyal, V. A gaussian process regression approach to predict the k-barrier coverage probability for intrusion detection in wireless sensor networks. *Expert Syst. Appl.* **172**, 114603 (2021).
103. Singh, A., Amutha, J., Nagar, J., Sharma, S. & Lee, C.-C. Lt-fs-id: Log-transformed feature learning and feature-scaling-based machine learning algorithms to predict the k-barriers for intrusion detection using wireless sensor network. *Sensors* **22**, 1070 (2022).
104. Singh, A., Amutha, J., Nagar, J., Sharma, S. & Lee, C.-C. Automl-id: Automated machine learning model for intrusion detection using wireless sensor network. *Sci. Rep.* **12**, 9074. https://doi.org/10.1038/s41598-022-13061-z (2022).
105. Singh, A., Amutha, J., Nagar, J. & Sharma, S. A deep learning approach to predict the number of k-barriers for intrusion detection over a circular region using wireless sensor networks. *Expert Syst. Appl.* **211**, 118588 (2023).
106. Singh, A. *et al.* Leveraging hybrid machine learning and data fusion for accurate mapping of malaria cases using meteorological variables in western India. *Intell. Syst. Appl.* **2023**, 200164 (2023).
107. He, X., Zhao, K. & Chu, X. Automl: A survey of the state-of-the-art. *Knowl.-Based Syst.* **212**, 106622 (2021).
108. Williams, R. Matrix-vector multiplication in sub-quadratic time: (some preprocessing required). *SODA* **7**, 995–1001 (2007).
109. Khan, M. R., Michael, H. A., Bresnyan, E. W. & Yu, W. Impacts of basin-wide irrigation pumping on dry-period stream baseflow in an alluvial aquifer in the kosi fan region of india and nepal. *Hydrogeol. J.* **30**, 1899–1910 (2022).

Acknowledgements

Author contributions

A.S.: methodology, conceptualization, validation, software, formal analysis, writing—original draft preparation, visualization, writing—review editing. K.G.: methodology, conceptualization, investigation, software, resources, writing—original draft preparation, writing–review editing, visualization, supervision.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.