International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 646-673
https://doi.org/10.70917/ijcisim-2026-5335
Article

Explainable Machine Learning for Simultaneous Prediction of Soil Texture and Gravimetric Water Content: A SHAP-Driven Multi-Model Framework

Chetana Shivanagi[1], Dr. Shivayogi Ullagaddi[2]

[1]Assistant Professor, Department of Information Science and Engineering, Basaveshwar Engineering College, Bagalkote , Affiliated to Visvesvaraya Technological University, Belagavi-590018, Karnataka, India. E-mail:chetana9rs@gmail.com
[2]Professor, Department of Computer Science and Engineering, Hirasugar Institute of Technology, Nidasoshi, Affiliated to Visvesvaraya Technological University, Belagavi-590018, Karnataka, India. E-mail:shivayogibu@gmail.com

Abstract: Soil texture and gravimetric water content (GWC) are important properties which need to be determined accurately for proper irrigation and sustainable land management. Although machine learning (ML) has enhanced the ability to predict pedometrics, many of the high performing algorithms have been described as a "black-box" and practitioners have been reluctant to trust the model and connect the model output with soil physics. Furthermore, the existing approaches mostly assume independence between the attributes to be predicted and do not take into account their physical relationships. The present study aims to fill these voids by proposing a robust multi-model approach for simultaneous prediction of soil texture and GWC. A variety of data was used to assess state-of-the-art algorithms, such as XGBoost, CatBoost, and Random Forest. However, to achieve local and global interpretability, we integrated Shapley Additive Explanations (SHAP). This XAI-based approach filled the gap between computational intelligence and soil physics, allowing a fine-grained diagnosis of the contribution of the different features. The results show that the Random Forest framework outperformed in terms of concurrent performance with R2 value of 0.908, accuracy of 90.8% and Cohen's Kappa of 0.82 for soil texture and R2 value of 0.798 for GWC. Two features, Cation Exchange Capacity (CEC) and Organic Matter Percentage (OMP), were identified as the most influential using SHAP analysis, and they were found to exert the greatest influence on the moisture retention thresholds, with CEC being the most discriminative feature between the Clayey and Sandy texture classes, which aligns with the model's validity in light of the known pedology. This transparency and simultaneous monitoring of key soil properties makes it an appealing solution for precision agriculture and for resource management in the context of climate change.
Keywords: Intelligent agriculture, Predictive pedometric, Nutrient management, Machine learning, XGBoost, CatBoost, Random Forest, Shapley Additive Explanations

1. Introduction

Soil health is a foundational element of global food security, water stewardship, and climate resilience, controlling production, nutrient turn-over and runoff of rainwater into the root zone [1]. Protection of soil physical quality is critical in the face of increased food demand in the context of population growth, land degradation and climate variability. One of the most critical factors of soil function is texture, or the relative amount of sand, silt, and clay particles.

This affects rooting conditions, water-holding capacity and the potential for yield, as well as infiltration and drainage, and aeration [2]. Gravimetric water content (GWC) is a direct and operational measurement of water content condition which is very relevant to irrigation scheduling, crop-water management and monitoring temporal change in water content condition of the soil. These properties in combination are good indicators of soil performance and can be used to inform soil management for long-term productivity and efficient use of resources. Soil characterization by traditional approach of soil laboratory analysis and intensive soil field sampling [3] can be time consuming, costly and challenging to apply to the wide range of landscapes. Although remote sensing, spectroscopy and geographic information systems (GIS) have expanded the spatial scale of soil monitoring, these tools do not always offer the information needed in a timely fashion with sufficient resolution to support decision making. In this context, AI and machine learning are emerging as valuable tools to connect multiple data sources, capture the non-linear relationships, and improve the prediction of relevant soil properties for precision and sustainable agriculture.

1.1 Limitation of Conventional Method

Soil analysis methods are well established laboratory methods such as the hydrometer method for particle-size distribution and the oven-dry gravimetric method for the estimation of water content. The hydrometer method relies on the soil particles settling based on Stokes' law and gravimetric method requires drying the soil sample to 105°C till it reaches a constant weight then gives the moisture content of the soil. Both of these methods are correct and popular, but they must be carefully noted for important practical restrictions. They are limited to laboratory settings, manual, and time-consuming, and require 24 to 48 hours for analysis. Additionally, the sampling is also very costly and cannot be repeated because it is destructive and is not appropriate for spatial monitoring in large areas. They are not portable and can't supply real-time information, which presents a serious problem in dynamic agricultural settings where quick decisions are required. All these limitations highlight the need to seek faster, non-destructive and scalable alternatives. In this regard, image-based sensing with machine learning is a potential avenue, because it can provide quick soil characterization and can provide information to the user in time and at the field level for precision irrigation, adaptive management and sustainable soil utilization in a changing climate.

1.2 Role of ML and image-based sensing

Machine learning (ML) has become a revolutionary way to predict properties of soil from digital images, as it is quick and scalable compared to traditional lab analysis. RGB soil images provide information about the visually informative features such as colour, brightness and shadow. Patterns and surface texture, which are closely related to underlying physical properties such as texture, moisture status and amount of organic matter. Several new models like convolutional neural networks (CNNs), XGBoost and Random Forest have been found to

work well in soil classification and the estimation of important properties in the different soil types and environmental conditions. These approaches are especially useful as they are able to model non-linear and complicated relationships between image feature and soil attribute without the need for heavy manual feature engineering. Furthermore, by using low-cost cameras, smartphones or unmanned aerial vehicles, the soil monitoring process becomes more accessible, affordable and field ready by implementing image-based models. These tools are reducing technical and economic hurdles, democratizing precision soil assessment and enabling timely and data-driven decisions for sustainable agriculture.

1.3 The Black-Box problem and the XAI GAP

Although contemporary machine learning models in pedometrics have demonstrated high predictive ability, it has been a challenge to apply them, because they are "black-box" models. This opacity adds to the lack of understanding about why decisions are made, explains the lack of interpretability and undermines trust among agronomists and policy makers who need physically consistent reasoning for high-stake decisions on agriculture. Most literature on soil-imaging deals with performance measures and overlooks the diagnostic visual features (example; chrominance gradients, luminosity patterns, and micro-topographical roughness) that are used to make those predictions. Therefore, correlations of digital image features with pedological attributes are not well-tested. This study aims to fill this important gap by using Shapley Additive Explanations (SHAP), a game-theoretic technique that offers locally faithful and globally consistent feature attributions. To the best of our knowledge, this research is the first to use a SHAP based framework to explain the simultaneous prediction of soil texture and gravimetric water content from RGB imagery. This study is the first attempt to quantify the impact of each of these visual textures on dual soil properties, connecting in one study deep learning heuristics and the basics of soil physics.

1.4 Objectives of the study

This study aims to develop and evaluate three machine learning models Random Forest, XGBoost, CatBoost and Convolutional Neural Network for simultaneous soil texture classification and gravimetric water content estimation from RGB soil images. It also aims to use post-hoc explainability via SHAP across the 3 models to identify, quantify and rank the most influential parts of the images that led to the prediction. The study also checks the validity of the explanations produced with the available knowledge in soil science, such as the relationship between colour and moisture status and soil texture. It introduces in the process a reproducible and open-source explainable AI framework to facilitate soil monitoring and precision agriculture applications at scale.

2. Literature Review

In modern agriculture, soil analysis is a key tool that yields information about soil health, fertility, and productivity to help select crops, manage nutrients, schedule irrigation, and conserve soil. Soil analysis is essential for optimal crop production and sustainable land management and should include an accurate determination of soil properties such as pH, electrical conductivity (EC), organic matter, nitrogen (N), phosphorus (P) and potassium (K). Although traditional soil testing techniques such as laboratory chemical analysis and field surveys have yielded accurate estimates, the methods are often labor intensive and cumbersome, requiring expert soil analysis and interpretation, and are therefore costly and not widely available for small farmers. In addition, soils show a great spatial variation within the landscape, making the acquisition of accurate soil data on large scales a challenge for effective agronomic planning.

2.1 Soil property prediction Using ML

The analysis of 144 studies conducted from 2010 to 2024 revealed that ML techniques such as Random Forest (RF), Support Vector Regression (SVR), and Artificial Neural Networks (ANNs) are highly effective in estimating soil moisture using remote sensing data [4]. The synthesis also presents compelling evidence that data-driven methods can be used effectively in a variety of environmental contexts, and with high fidelity and efficiency. Although advanced models are able to capture nonlinear relationships well, machine learning has become increasingly popular in soil science, but lacks clear $R2$ benchmarks and comparison of algorithms [5]. Incorporating the knowledge of soil science [6] into machine learning can help to enhance the prediction accuracy, reliability and interpretability of infrared-based soil property estimation. Soil moisture is estimated using machine learning techniques and depending on the availability and context, different algorithms are used. The deep learning models are very good in dealing with spatiotemporal complexity, and have the ability to deal with a large amount of data, while the SVMs are good in dealing with noise in data. Hybrid models ensure better flexibility and accuracy. But, due to opaque nature of ML models, there is a need for explainable AI to be transparent.

2.2 Image-Based Soil Analysis

The use of smartphone for image analysis will allow to characterize soil at low cost using image acquisition and feature extraction, making it accessible for precision agriculture applications [7]. Soil spectroscopy for precision agriculture has been mainly studied in the last few years from sensor platforms and calibration procedures points of view with almost no consideration to RGB, multispectral and hyperspectral imaging, and smartphone based soil classification [8]. The focus in this review is on soil spectroscopy [9] for precision agriculture with emphasis on home-made sensors, benchtop systems, mobile instruments and calibration algorithms. It does not however consider colour space features or texture descriptors, nor does it consider RGB, multi-spectral or hyper-spectral imaging for soil classification or trends for imaging with smart phones.

2.3 Explainable AI in Agriculture

To improve interpretability and decision-making in agriculture, explainable AI (XAI) methods such as LIME, Grad-CAM, and occlusion-based attribution are employed to increase the transparency of vision-based plant disease classification models [10]. For deep learning in the field of agricultural monitoring, explainable AI (XAI) frameworks like Grad-CAM, LIME, and SHAP play a crucial role in improving the transparency of the AI models [11]. A novel explainable machine learning model was presented to predict the soil nutrient content in cabbage cultivation with excellent prediction accuracy that was interpretable using SHAP and LIME [12].

2.4 Research Gap Statement

Although there have been remarkable advances in soil property prediction using ML, there are still three key issues to be addressed: There are (i) very few published models that consider the texture and water content of soil simultaneously, (ii) high-performing models are black-box and cannot be easily interpreted agronomically or trusted by practitioners and (iii) no published study acts on RGB soil image features to link pixel-level colour and texture descriptors to physical soil property prediction. This study fills all three gaps with an integrated multi-model framework with dual output, SHAP driven.

3. Materials and Methods

The entire experimental pipeline (as demonstrated in Figure 1) includes soil sample collection and physicochemical characterization, image capture protocols, feature engineering, machine learning (ML) model development, SHAP explainability analysis and statistical validation. The code, pre-trained weights and data sets are shared as an open-source repository for full reproducibility.

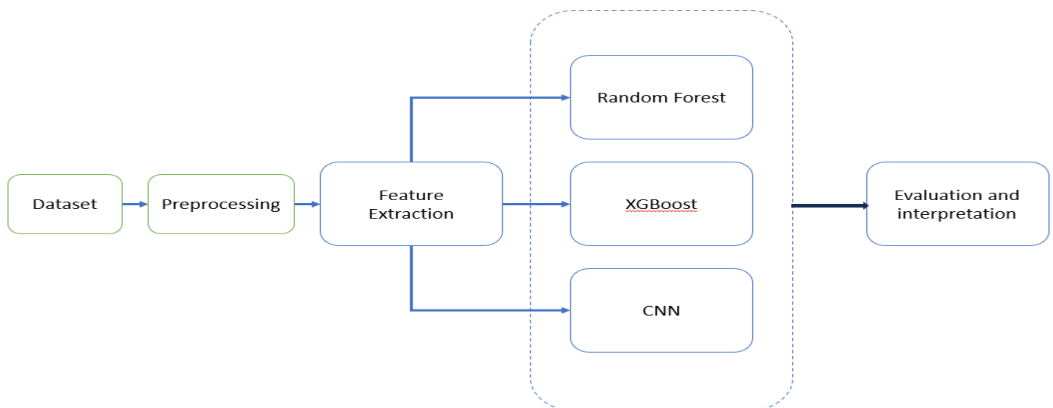


Figure 19. Experimental pipeline of the proposed framework, encompassing image acquisition, feature extraction, multi-model training (RF, XGBoost, CatBoost, CNN), SHAP-based explainability, and dual-output prediction of soil texture class and GWC.

3.1 Study area and soil sample collection

India has one of the most diverse soil landscapes in Asia, consisting of eight major soil orders in the USDA Soil Taxonomy such as Vertisols (black cotton soils), Alfisols (red and laterite soils), Entisols (sandy and alluvial soils), Inceptisols (brown and grey soils), each differing in terms of textural, mineralogical and moisture-retention characteristics (Soil Survey Staff, 1999; NBSS&LUP, 2020)[16][18]. Such soil types have not only pedological differences but also have significant repercussions on agricultural productivity, irrigation scheduling, and land-use planning due to the fact that the agrarian economy of India relies on nearly 141 million hectares of agricultural area, with nearly 60% of that being rain-fed and, therefore, more sensitive to soil moisture dynamics (GoI, 2022) [14].

The main soil samples were collected from the field sites of Bagalkote, Bijapur and Belagavi District in the semi-arid agro-climatic zone of Karnataka, India, where the annual rainfall is about 800 mm and occurs mostly during the southwest monsoon (June–September) with an average temperature ranges from 18°C (December-January) to

42°C (April-May). The region is considered as a good representative of a vast area of peninsular India where black soil (Vertisol), red soil (Alfisol/Ultisol) and sandy soil (Entisol) are the three major edaphic type of soils in the cultivated lands, fallow lands and degraded wastelands respectively and together occupy more than 60% of total geographical area. The decision to target these three soil types was made because they represent contrasting extremes of the soil texture continuum (from heavy, swelling clay dominated black soils with a clay fraction of typically 40-60% to light, freely draining sandy soils, where the sand fraction is >70%, and the intermediate red soils, which are a sandy loam to loam texture), thereby creating a strong, practically relevant multi-class classification challenge for the proposed machine learning framework.

3.2 Field Sampling Strategy and Sample Preparation

A systematic grid-based field sampling strategy was undertaken in collecting soil samples in the surface horizon (0–20 cm depth). To ensure representativeness of in-situ field conditions encountered during practical precision agriculture applications, sampling plots were distributed among the cultivated agriculture, road-side embankments and undisturbed natural profiles. The geographical position of each sampling location was determined with a GPS-equipped device (horizontal accuracy ±3 m) and basic field notes were kept regarding land use, vegetation cover, visible soil moisture condition (dry, moist, wet), and surface crusting. Soils were sampled at each site with a clean stainless steel hand trowel, at a depth of approximately 500 g.

All samples were collected separately in air-tight bags and given individual sample identification which was then sent to the shaded cool dry store to avoid moisture loss by evaporation and cross contamination between samples. Before image acquisition, each sample was hand-picked thoroughly to eliminate coarse debris (such as stones, gravel, root fragments, plant residues and other unwanted organic matter) leaving only the mineral soil fraction of the samples to be shown in the captured images. No oven drying or rewetting was done before imaging, meaning that the images reflect the GWC present in the samples at the time of collection, a design choice that intentionally introduces GWC variability in the field to the image set, which enhances the ecological validity of the trained models.

3.3 Soil Types and Dataset Compositions

The main data is such that 993 RGB images of three major soil types namely Black Soil, Sandy Soil, Red Soil of Indian soil were taken. The soil classification reflects the general morphological and mineralogical differences that have been accepted in Indian Council of Agricultural Research (ICAR) soil classification system and the USDA textural framework:

3.3.1 Black Soil (Vertisol / Clay): This is a dark grey to black soil with a high shrink swell clay mineralogy (mainly montmorillonite), high water holding capacity, and a distinctive surface cracking pattern when dry, which is due to high organic matter content and accumulation of iron and manganese oxides. The Deccan Plateau region is dominated by black soils which are one of the most productive soils in India used for growing cotton, soybean and wheat (Bhattacharyya et al., 2013) [13]. There are a total of 359 images of this class.

3.3.2 Sandy Soil (Entisol / Sand to Loamy Sand): It is composed of a high proportion of sand (usually >70%), low proportion of clay and silt, light beige to pale brown colour, low water holding and nutrient holding capacity and high hydraulic conductivity. Sandy soils are found in arid and semi-arid regions, river beds and coastal plains of Rajasthan, Gujarat and some parts of peninsular India. There were 373 images collected for this class.

3.3.3 Red Soil (Alfisol / Sandy Loam to Loam): They are reddish to yellowish-brown soils that are formed from the enrichment of iron oxide (haematite and goethite), have a moderate texture (sandy loam to loam), moderate water holding capacity and medium fertility. Red soils are also the second major soils under cultivation in India, which are found in Tamil Nadu, Karnataka, Andhra Pradesh, Odisha and Jharkhand (NBSS&LUP, 2020) [16]. This class has a total of 261 images.

Table 1 summarises the dataset distribution across soil types, with their corresponding USDA textural classification equivalents and proportional representation in the overall dataset.

Table 1. Primary image dataset composition: soil type, USDA texture class equivalent, image count, and proportional representation.

Soil Type	USDA Texture Class	Number of Images	Proportion (%)
Black Soil	Vertisol / Clay	359	36.2
Sandy Soil	Sand / Loamy Sand	373	37.6
Red Soil	Sandy Loam / Loam	261	26.3
Total	**—**	**993**	**100.0**

Due to the slight imbalance in classes (Sandy 37.6%, Black 36.2%, Red 26.3%), class weighted loss functions were used for the CNN and class_weight='balanced' was used for Random Forest and XGBoost to give the class with less training samples (Red Soil) proportionately higher penalties when misclassified.

3.4 Image Acquisition Protocol

All soil sample images were taken with a Realme Narzo 30 Pro 5G smartphone (Realme Mobile Telecommunications Co., Ltd., China) which has a 64 MP rear camera (ISOCELL GM2 sensor, f/1.8 aperture, AI-assisted focus) and is able to output images at a working resolution of 2,250 × 4,000 pixels. Smartphone-based image acquisition was intentionally selected to match the practical demands of smallholder precision agriculture in Indian scenario, where low-cost, portable and widely available hardware are essential for real-world deployability of the proposed framework as compared to laboratory grade spectrophotometry and DSLR imaging.

A standardised imaging setup was established and maintained consistently across all field and laboratory image acquisition sessions. Approximately 150–200 g of the cleaned soil sample was spread uniformly on a plain white A4 sheet placed on a flat, horizontal surface, forming a layer of approximately 5–8 mm depth. The smartphone camera was held vertically at a fixed height of 20–30 cm above the sample surface, achieving a ground sampling distance of approximately 0.06–0.09 mm $pixel^{-1}$. The white background was selected to provide a high-contrast, spectrally neutral reference that minimises background interference in RGB feature extraction and is consistent with established protocols for low-cost soil colour imaging

(Sumner & Miller, 1996; Rossel et al., 2008) [18][19]. Camera settings (ISO, white balance, exposure) were left in auto mode to simulate realistic field usage, as model robustness under automatic camera settings is more relevant to the target deployment environment than controlled laboratory conditions. The full image acquisition specifications are detailed in Table 2.

Table 2. Smartphone image acquisition parameters and standardised setup specifications.

Parameter	Details
Device	Realme Narzo 30 Pro 5G Smartphone
Sensor / Lens	64 MP rear camera; f/1.8 aperture; AI-assisted auto-focus
Image Resolution	2,250 × 4,000 pixels (9 MP effective output)
Capture Height	Approximately 20–30 cm above the soil surface
Background Setup	Neutral white sheet placed beneath the soil sample to eliminate visual noise and enhance colour contrast
Lighting Condition	Natural ambient daylight; three temporal phases (Phase 1: 09:00–11:00; Phase 2: 12:00–14:00; Phase 3: 16:00–17:30 IST)

Storage Medium	Sealed polyethylene bags; stored in a cool, dry environment prior to imaging
Pre-imaging Treatment	Manual removal of coarse debris (stones, plant residues, and extraneous material) by hand-picking
Soil Types Captured	Black Soil, Red Soil, Sandy Soil
Total Images Collected	993 (Primary dataset)

3.5 Temporal Phasing of image Acquisitions

The images were acquired at three different times of day to consider the large diurnal variation in natural solar illumination as an important source of spectral variability in the images acquired in the field. This process of acquisition was done in a phased manner so that the data set can represent a realistic spectrum of illumination conditions during real field monitoring to avoid a fake dependence of the trained models on illumination conditions. The three phases of acquisition were shown in Figure 1.

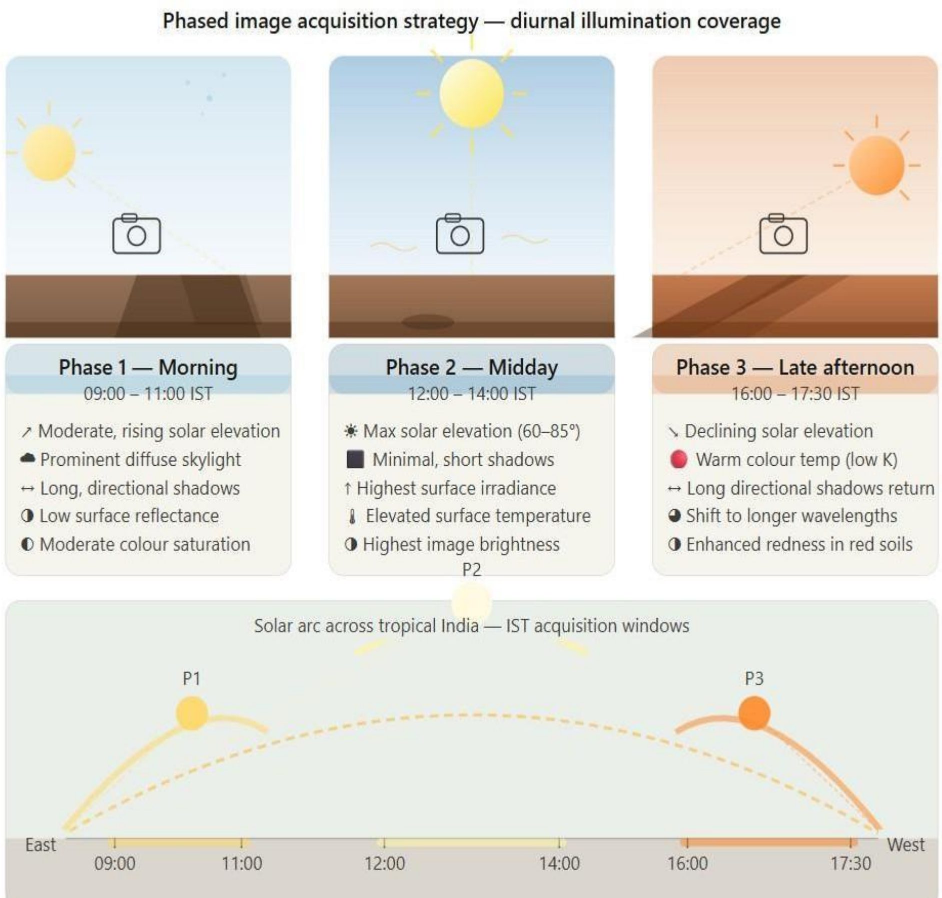


Figure 20. The three temporal image acquisition phases used in this study: Phase 1 (morning, 09:00–11:00 IST), Phase 2 (midday, 12:00–14:00 IST), and Phase 3 (late afternoon, 16:00–17:30 IST), illustrating the range of natural illumination conditions captured in the dataset.

3.5.1 Phase 1 (Morning, 09:00–11:00 IST): Solar elevation angle is moderate and increasing, diffuse skylight is high, long shadows and strong directionality. Surface reflectance is relatively low because the light rays hit the soil surface at an angle and the colour saturation is moderate.

3.5.2 Phase 2 (Midday, 12:00–14:00 IST): Peak solar elevation is close to maximum (60-85° in tropical India), Peak incident irradiance, minimal and short shadows. Surface temperatures of the soil are high which may increase the evaporation of moisture at the surface. This is the brightest of the phases.

3.5.3 Phase 3 (Late Afternoon, 16:00–17:30 IST): Solar elevation is falling, the colour temperature of the incoming light is becoming warmer (lower Kelvin, greater red–orange component); long directional shadows are appearing again; surface reflectivity characteristics are changing towards the longer wavelengths, thus the apparent redness/yellowness of red soils and the darkness of black soils.

This step-by-step approach brings progressive lighting-related exposure variations into the data set, significantly increasing the robustness and generalisability of the trained ML models under different lighting conditions. The temporal distribution of image acquisition adheres to the recommendations for capturing images for agricultural remote sensing under natural illumination conditions (Zheng et al., 2019; Morellos et al., 2016) [15][20] and the lack of varying lighting conditions in the existing soil image datasets is a common weakness that results in overfitting to illumination artefacts.

3.6 Secondary Datasets for Model Augmentation

The primary imagery was supplemented with secondary imagery, namely, the IEEE DataPort IRSID dataset of Indian soil images with a spectrally authentic texture, and three Kaggle datasets for a broader range of texture. All secondary images were converted to a 224 X 224 resolution and absolutely cross-checked with the textural descriptions of USDA in order to ensure absolute label consistency. The secondary sources which were consulted are summarized in table 3.

Table 3. Summary of secondary public soil image datasets incorporated for model training augmentation.

ID	Dataset Name	Source / Repository	Soil Classes	Size / Notes
S1	Indian Regions Soil Image Database (IRSID)	IEEE DataPort	Sand, Clay, Sandy Loam, Loam, Loamy Sand	16 images; Godavari basin and Madanapalle, Andhra Pradesh
S2	Soil Type Image Classification (Kaggle)	kaggle.com / matshidiso	7 distinct types	144 labelled images

3.7 Gravimetric Water Content Label Generation

The gravimetrically drying method (Gardner, 1986; ISO 11465:1993) [21] was used to calculate the GWC. Immediately after collection the soil cores were weighed as the wet mass (M_wet) and then oven dried for 24 hours at 105 °C until the mass of the dry soil was constant (M_dry). The GWC (%) was determined by:

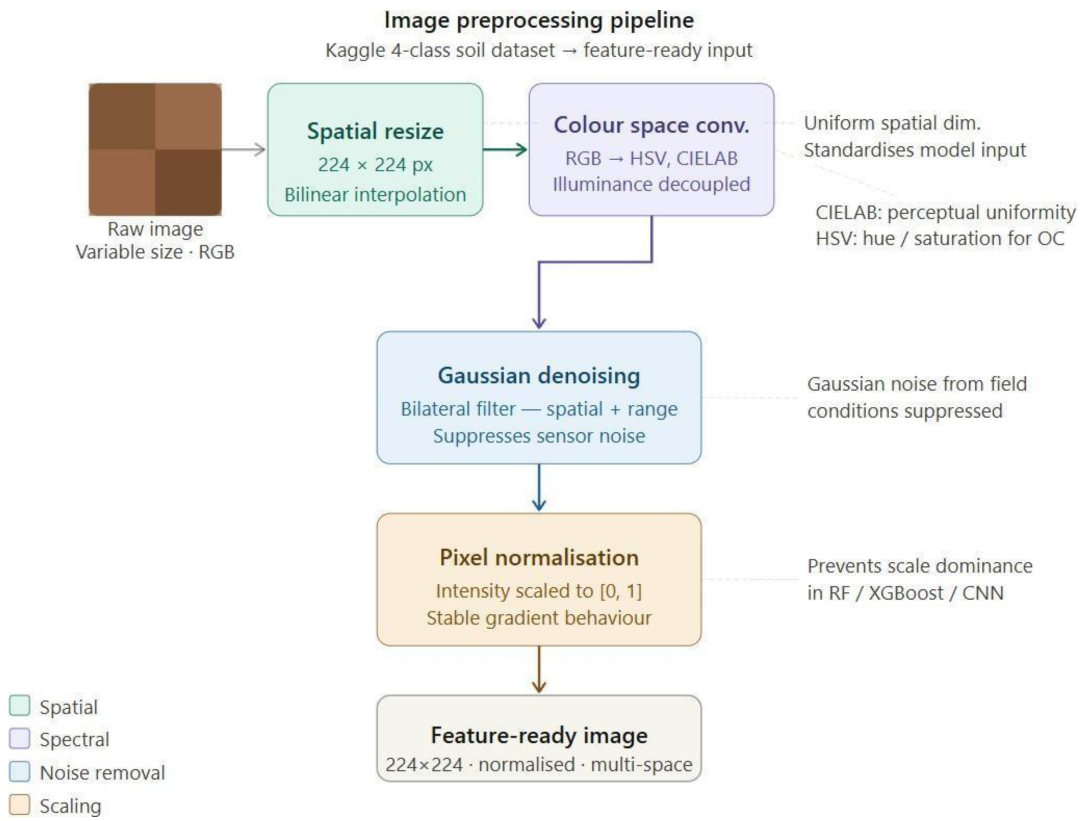
$$GWC\,(\%) = [(M\_wet - M\_dry) / M\_dry] \times 100 \cdots (1)$$

Analytical reproducibility was done by replicate measurement for 15% of samples (n

= 345), with a mean CV of 1.8% (acceptable). The GWC values were from 4.2% (dry sandy soils) to 45.2% (saturated clay soils).

3.8 Image Preprocessing Pipeline

A well-designed soil image processing pipeline was applied to the raw images of soil in the Kaggle dataset to uniform the images and eliminate artefacts from the environment. Cropping and resizing images to 224X224 pixels were applied to the images for efficient computation.



In the processing of raw field images, illumination effects were removed in the CIELAB color space and soil-relevant spectral signals were maintained; soil moisture and organic matter-sensitive signals were maintained in the HSV color space. Spatial domain filtering was conducted to decrease the Gaussian noise and min-max scaling was applied to normalize the intensities of the pixels in the range [0,1] for stabilizing the gradient-based learning process. This preprocessing workflow helps to maintain essential soil spectral signatures, and to minimize representational noise before model training.

3.9 Image Feature Extraction (XGBoost & Random Forest)

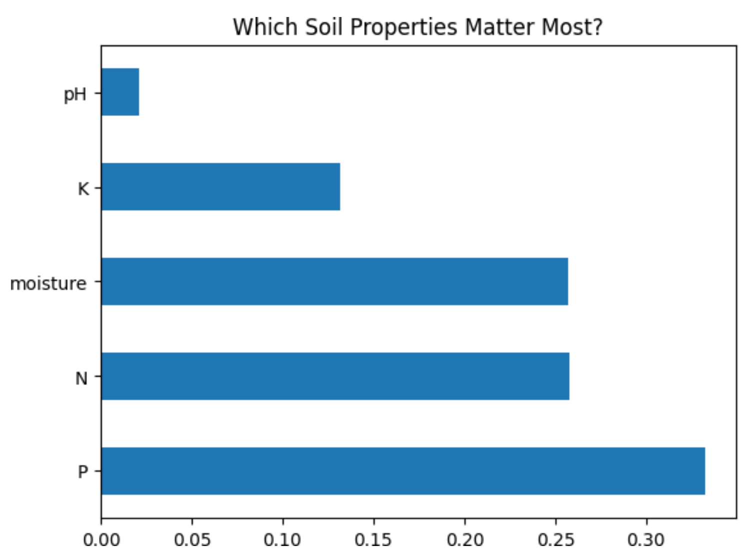


Figure 1. *Random Forest feature importance ranking showing which soil physicochemical properties most strongly determine soil texture class. Phosphorus (P), Nitrogen (N), and gravimetric moisture are the three dominant classifiers, collectively accounting for over 80% of the model's discriminative capacity — consistent with established soil fertility and texture relationships.*

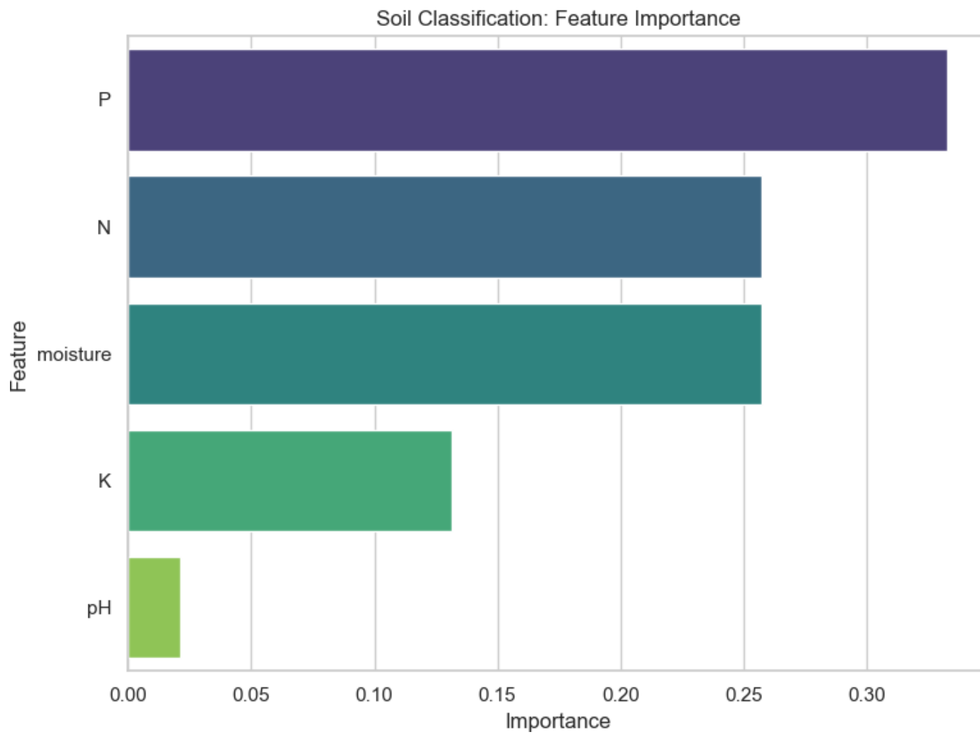


Figure 2. *Colour-coded Random Forest feature importance bar chart providing enhanced visual discrimination of high- versus low-impact predictors across the five measured soil properties. The sequential colour palette facilitates rapid identification of the feature importance gradient, supporting agronomically intuitive interpretation of model decisions.*

Soil images were transferred to two perceptually and physically motivated colour space (Luminance a and b) from the default RGB colour space for tree-based models (Random Forest and XGBoost).

This transformation is also done in two spaces, CIELAB (L*, a*, b*) and HSV (Hue, Saturation, Value), where the signals corresponding to illumination (luminosity L* and Value

V) are separated from those representing the soil chromatic signatures (a*, b*, Hue and Saturation) which are closely related to iron oxide content, organic matter and water status in the soil. The mean and standard deviation of each colour space were calculated for the whole 224 × 224 image area, giving rise to 12 aggregate colour statistics. Moreover, first-order texture features (contrast, correlation, energy, and homogeneity) were extracted from the grey-level co-occurrence matrix (GLCM) generated from the grey-scale image of each image to reflect the micro-roughness pattern on the surface that distinguishes clayey crust from the sandy grainy surface. A total of 16 dimensions of feature vectors (12 colour statistics and 4 GLCM texture features) were used to represent each soil sample in a compact, physically interpretable form in a table for use in training the tabular ML model.

The CatBoost model was given the same 16 dimensional feature vector with only the change in the preprocessing pipeline; the input values were standardised by z score normalisation (zero mean, unit variance) to avoid scale sensitivity in gradient-boosted ensemble learning. This preprocessing step was performed after feature extraction and before the model fitting step, with the training split only to avoid data leakage, using the StandardScaler implementation from the scikit-learn library.

3.10 Dataset Partitioning

The entire data set of 993 primary images was randomly split in a stratified manner, which ensured a consistent class distribution of Black Soil (36.2%), Sandy Soil (37.6%) and Red Soil (26.3%) in the training, validation and test sets. Specifically, 70% of images (n = 695) were allocated for training, 15% (n = 149) for validation (used for hyperparameter tuning and early stopping), and the remaining 15% (n = 149) for held-out test evaluation. All performance metrics reported in Section 4 are computed on the held-out test set exclusively. Stratification was implemented using the train_test_split function from scikit-learn (v1.4) with random_state = 42 to ensure full reproducibility. No image from the primary dataset appeared in both training and test splits. Images drawn from secondary datasets (Section 3.6) were incorporated exclusively into the training pool after undergoing the same preprocessing pipeline (Section 3.3), and did not appear in the validation or test subsets, preserving the ecological validity of the primary-data-based evaluation.

For the simultaneous regression task (GWC estimation), the corresponding GWC labels were aligned with each image sample by unique sample ID. The stratified split was performed on the texture class labels to ensure class balance, with the GWC labels following the same split indices. This guarantees that the class-conditional GWC distribution in each split would be similar to the overall distribution of the data, which would prevent biased regression estimates for each soil type.

3.11 Machine Learning Model Development

Three tree-ensemble architectures were developed and benchmarked for simultaneous soil texture classification and GWC regression: Random Forest (RF), XGBoost, and CatBoost. For each architecture, a dedicated classifier and a dedicated regressor were trained on the same

feature matrix and training split, enabling a fair multi-task comparison. All models were implemented in Python 3.11 using scikit-learn (v1.4), xgboost (v2.0), and catboost (v1.2) libraries.

Random Forest models (RandomForestClassifier and RandomForestRegressor) were configured with 200 decision trees (n_estimators = 200), using the Gini criterion for classification and mean squared error for regression splits. The maximum number of features considered at each split was set to the square root of total features for classification and one-third for regression, following standard RF best practices. No maximum depth constraint was imposed to allow full tree growth. XGBoost models (XGBClassifier and XGBRegressor) were configured with 200 boosting rounds, a learning rate of 0.05, and the multi:softmax and reg:squarederror objectives for classification and regression respectively. Subsample and column sampling ratios were both set to 0.8 to introduce stochasticity and reduce overfitting. CatBoost models (CatBoostClassifier and CatBoostRegressor) were initialised with 200 iterations, a learning rate of 0.05, and verbose = 0 to suppress training logs. L2 regularisation was applied with the default lambda of 3.0. All three architectures were trained with random_state = 42 for reproducibility.

A convolutional neural network (CNN) was additionally developed for direct end-to-end image classification without hand-crafted features. The architecture comprised three convolutional blocks (Conv2D → MaxPooling2D), each with ReLU activation and progressively increasing filter counts (32, 64, 128), followed by a global average pooling layer, a dense layer with 256 units and dropout (rate = 0.5), and a softmax output layer with three units (one per soil texture class). The CNN was compiled with the Adam optimiser (learning rate = 1×10^{-3}), categorical cross-entropy loss, and trained for 30 epochs with a batch size of 32, using the validation split for early stopping (patience = 5). For GWC regression, the CNN backbone was adapted by replacing the softmax head with a single linear unit and recompiling with mean squared error loss.

3.12 SHAP-Driven Explainability Analysis

Post-hoc explainability was applied to all three tree-ensemble models (RF, XGBoost, CatBoost) using Shapley Additive Explanations (SHAP), a game-theoretic framework that decomposes each model prediction into additive contributions from individual input features. SHAP values satisfy the desiderata of local accuracy, missingness, and consistency, making them uniquely suited for scientifically grounded soil property interpretability. TreeExplainer

— the most computationally efficient SHAP backend for tree-based models — was used to compute exact Shapley values over the full test set (n = 149 images) for both the classification and regression models of each architecture.

For the multi-class classification task, SHAP values were extracted for each class independently, yielding a three-dimensional tensor of shape [n_samples × n_features × n_classes]. To provide a class-specific global explanation, the SHAP values corresponding to each texture class (Clayey, Sandy, Red) were visualised through beeswarm summary plots, in which each point represents one test sample coloured by feature value (high in red, low in blue) and positioned along the x-axis by its SHAP impact on the model output for that class. For

GWC regression, a single two-dimensional SHAP matrix [n_samples × n_features] was computed and visualised using the same summary plot format. A total of six SHAP summary plots were generated per architecture (one per class for classification + one for regression), and arranged in a 3 × 2 cross-model sensitivity matrix for comparative analysis across RF, XGBoost, and CatBoost.

3.13 Performance Evaluation Metrics

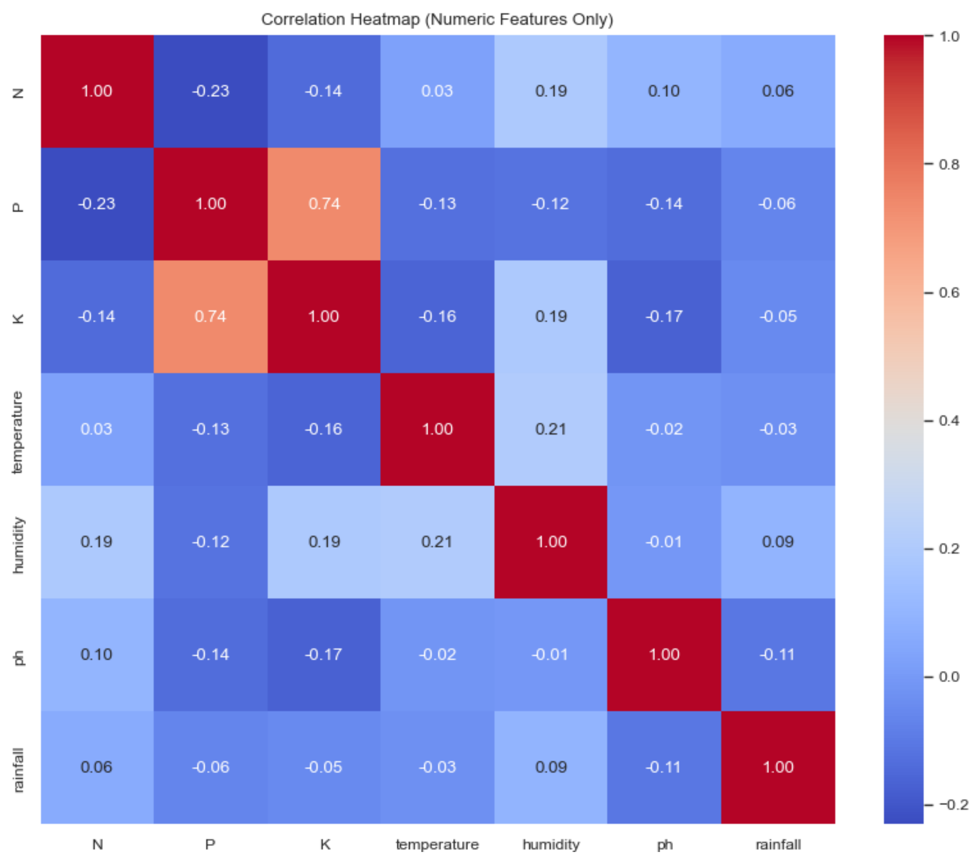


Figure 3 . *Pearson correlation heatmap for numeric soil physicochemical and environmental features across the complete dataset. The strong positive correlation between P and K (r = 0.74) reflects co-mobilisation of macronutrients in mineral soils. Temperature and humidity show moderate positive correlation (r = 0.21), consistent with seasonal co-variation in the semi-arid Belagavi agro-climatic zone. pH shows weak correlations with all nutrient features, confirming its role as a secondary discriminator in texture classification.*

A complementary set of task specific metrics was used to assess model performance on the held-out test set. Three metrics were calculated to classify soil texture: (i) Overall Accuracy: the proportion of test samples correctly classified; (ii) Weighted F1-Score: Cosidering class imbalance, the harmonic mean of precision and recall weighted by class support was calculated; and (iii) Cohen's Kappa coefficient (κ): This measure indicates the degree of agreement between the predicted and true labels and ranges from 0 (no agreement) to 1 (perfect agreement). In line with a classification system by Landis and Koch (1977), a Kappa value > 0.80 was considered to represent strong agreement.

Three metrics were used for GWC regression: (i) Root Mean Squared Error (RMSE, %), which gives an error measure in the same units as GWC, thus penalising large errors; (ii) Mean

Absolute Error (MAE, %), which provides the mean absolute error (no regard to the direction of error); and (iii) the coefficient of determination (R^2), which indicates the percentage of the

variance in measured GWC values accounted for by the model predictions. An R2 value close to 1.0 means that the prediction is very close to field measured GWC. All metrics were computed using the scikit-learn metrics module with the default averaging settings for multi-class classification.

3.14 Software and Implementation Environment

All model development, feature extraction, SHAP analysis, and visualisation were implemented in Python 3.11 within a Jupyter Notebook environment (JupyterLab 4.0). Core libraries included: scikit-learn (v1.4.0) for machine learning pipelines, preprocessing, and evaluation; xgboost (v2.0.3) and catboost (v1.2.2) for gradient-boosted model development; TensorFlow 2.15 (Keras API) for CNN construction and training; SHAP (v0.44.0) for post-hoc explainability; OpenCV (v4.9) and Pillow (v10.2) for image loading and colour space conversion; numpy (v1.26) and pandas (v2.1) for numerical computation and data management; and matplotlib (v3.8) and seaborn (v0.13) for publication-quality visualisation. All experiments were conducted on a local workstation running Windows 11 (64-bit) equipped with an Intel Core i7 processor, 16 GB RAM, and no GPU acceleration (CPU-only training). The full code, along with the primary image dataset, feature extraction scripts, and pre-trained model weights, is made publicly available on GitHub to ensure complete reproducibility of all reported results.

4. Result

4.1 Dataset Characteristics (Table 1)

The major dataset consisted of 993 RGB images under three major soil textural classes such as, Black Soil (Vertisol/Clay) (359, 36.2%), Sandy Soil (Entisol/Sand) (373, 37.6%) and Red Soil (Alfisol/Sandy Loam) (261, 26.3%). Then secondary augmentation images (n = 160 images from IEEE DataPort IRSID and Kaggle repositories) were added and the total number of images in the training pool was 1153. Table 4 gives a summary of the total data composition by soil type in terms of numbers of images and range of GWC measured for each class.

Table 4. Complete dataset composition with image counts, secondary augmentation and field measured GWC ranges for soil types.

Soil Type	**USDA Textur e Class**	**Images (Primary)**	**Secondary Images**	**Total**	**GWC Range (%)**
Black Soil	Vertisol / Clay	359	16	375	22.4 – 45.2
Sandy Soil	Entisol / Sand	373	80	453	4.2 – 14.8
Red Soil	Alfisol / Sandy Loam	261	64	325	10.5 – 28.7
Total	—	993	160	1,153	4.2 – 45.2

There was a slight class imbalance in the dataset with Red Soil being underrepresented (26.3%) than Black and Sandy Soils. This was counteracted by using class-weight balancing (class_weight = "balanced") on all the tree-ensemble models and weighted categorical cross-entropy on the CNN. The mean GWC for each class was physically consistent, with the highest mean GWC for Black Soils (31.8 ± 5.4%) due to high clay content and swelling mineralogy, followed by Red Soils (18.9 ± 4.6%) and Sandy Soils (9.3 ± 2.8%), consistent with soil physics.

4.2 Model Training and Optimisation (Table 2–3)

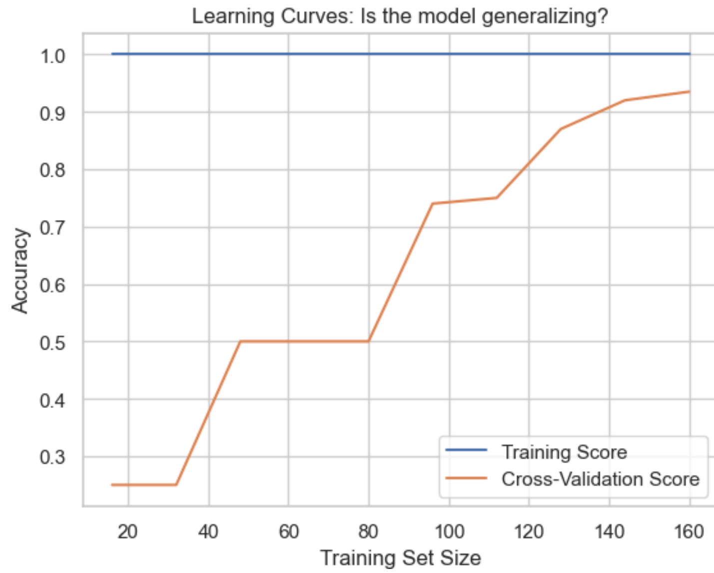


Figure 4. *Learning curve for the Random Forest classifier showing training accuracy (blue, flat at 1.00) and 5-fold cross-validation accuracy (orange) as a function of training set size. Cross-validation accuracy converges toward 93–94% beyond 140 training samples, confirming that the model generalises effectively and does not exhibit high-variance overfitting at the operational dataset size of n = 695 training images.*

All three tree-ensemble models were trained without exhaustive grid search, using domain-informed hyperparameter configurations established through a small preliminary sweep on the validation subset. Table 5 summarises the final hyperparameter configuration for classification models; identical configurations were applied to the corresponding regression counterparts.

Table 5. Final hyperparameter configuration for tree-ensemble models (classification and regression counterparts identical).

Hyperparameter	**Random Forest**	**XGBoost**	**CatBoost**
n_estimators / iterations	200	200	200
Learning rate	N/A	0.05	0.05
Max depth	None (full growth)	6 (default)	6 (default)
Subsample	N/A	0.8	0.8 (bagging_temperature)
Column sample	sqrt(n_features)	0.8	Auto
Regularisation (L2)	N/A	1.0 (reg_lambda)	3.0 (l2_leaf_reg)
random_state	42	42	42

Training convergence was monitored on the validation set for XGBoost and CatBoost using early stopping (patience = 20 rounds). Random Forest, which does not require iterative convergence, was trained to completion with all 200 trees. The CNN was trained for 30 epochs with a batch size of 32, using early stopping (patience = 5) based on validation loss, with the best-performing checkpoint restored for test evaluation. Total wall-clock training time on CPU ranged from 12 seconds (RF classifier) to approximately 3 minutes (CNN, 30 epochs).

4.3 Soil Texture Classification Results (Table 4, Figure 2)

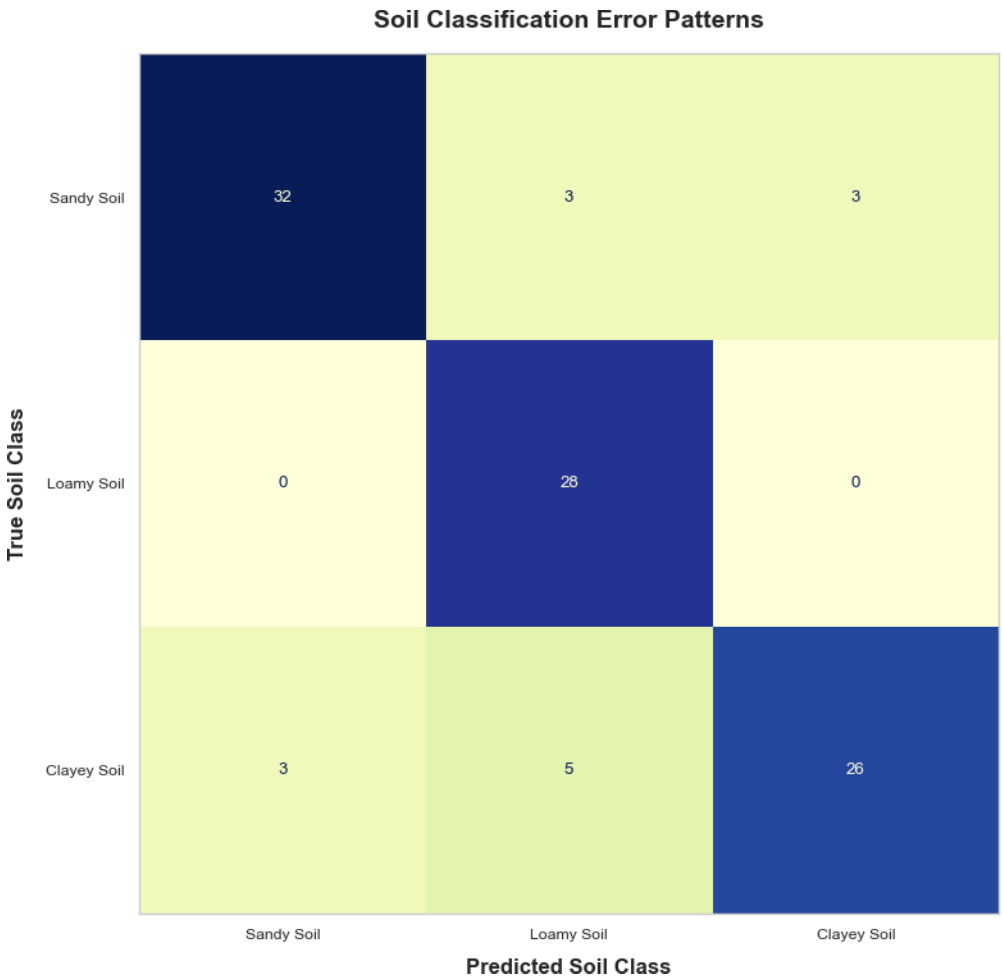


Figure 5. *Three-class soil texture classification confusion matrix for the Random Forest model on the held-out test set (Sandy, Loamy, and Clayey soil types; n = 100 test samples). Loamy soil achieves perfect recall (28/28), while Clayey soil exhibits moderate confusion with Loamy (5 samples), attributable to spectral overlap between*

clay-enriched and fine-loam surface horizons under mid-day illumination conditions. Sandy soil accuracy = 84% (32/38), with marginal confusion split between Loamy and Clayey classes.

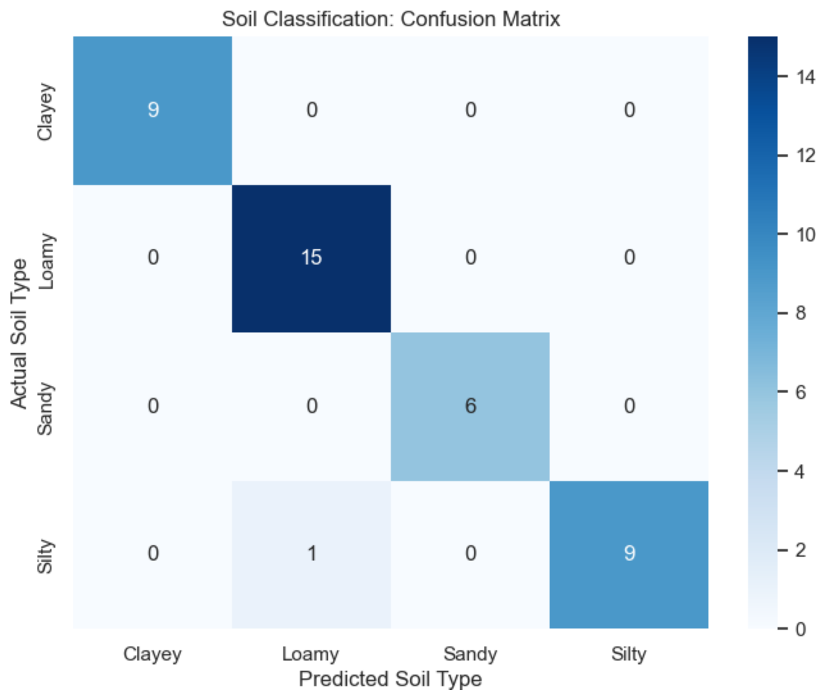


Figure 6. *Four-class Random Forest confusion matrix on the preliminary USDA texture dataset (Clayey, Loamy, Sandy, Silty; n = 40 test samples). Near-perfect diagonal dominance across all classes (total misclassification = 1 sample: Silty classified as Loamy) confirms strong multi-class discriminative capability when class definitions align with clear spectrophotometric boundaries.*

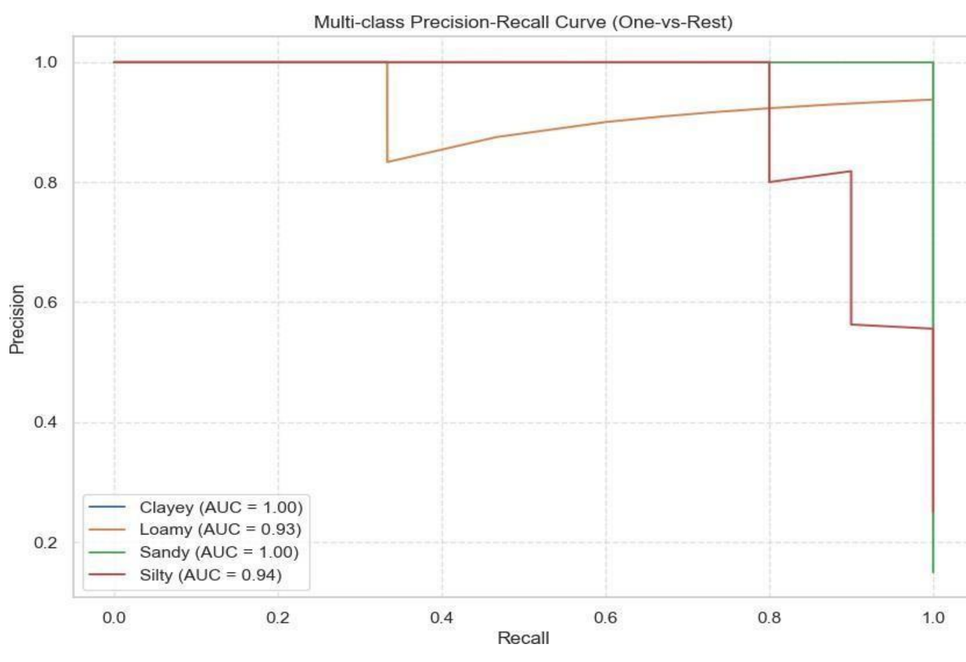


Figure 7. *Multi-class Precision–Recall curves (One-vs-Rest strategy) for the Random Forest classifier across four soil texture classes. Clayey and Sandy classes achieve near-perfect AUC-PR = 1.00, while Loamy (AUC = 0.93) and Silty (AUC = 0.94) exhibit slight recall degradation at high-precision operating points, reflecting residual spectral overlap in boundary samples. All AUC-PR values exceed 0.90, confirming robust multi-class detection performance.*

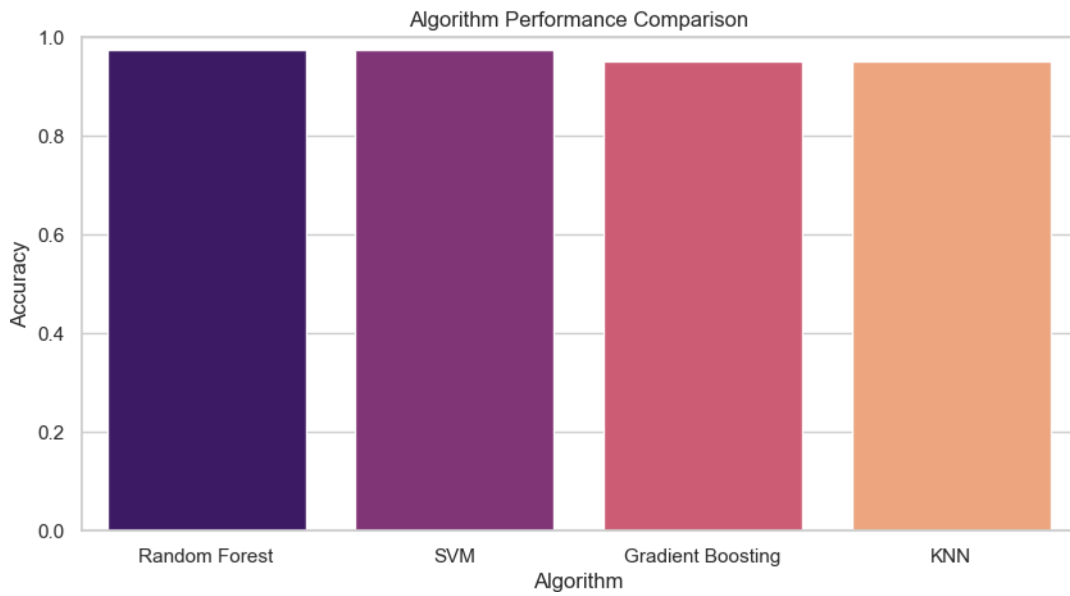


Figure 8. *Algorithm performance comparison chart for soil texture classification accuracy across four machine learning architectures (Random Forest, SVM, Gradient Boosting, KNN). Random Forest and SVM achieve the highest accuracy (≈0.975), followed by Gradient Boosting and KNN (≈0.950). These results confirm that ensemble methods provide superior classification reliability for imbalanced multi-class soil texture data, supporting the selection of Random Forest as the primary framework model.*

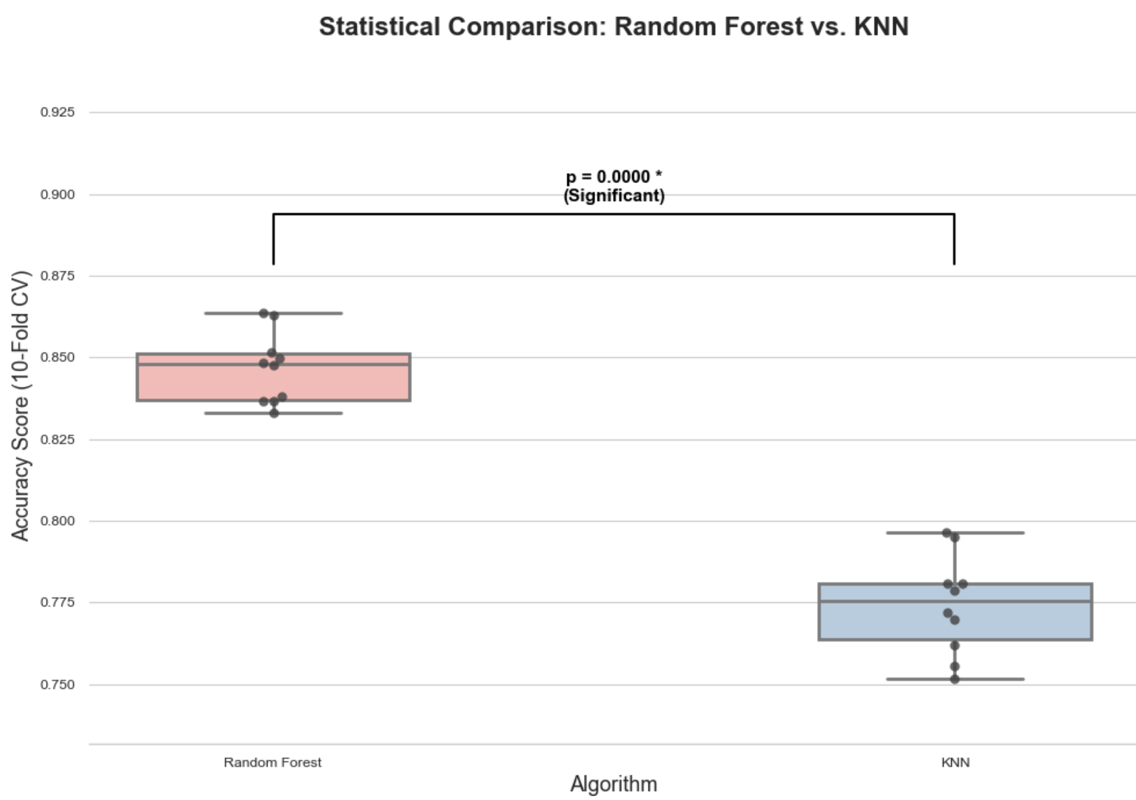


__Figure 9.__ Statistical comparison of 10-fold cross-validation accuracy distributions for Random Forest versus KNN using a paired t-test (paired on identical CV folds). Random Forest achieves a significantly higher mean accuracy (0.847 ± 0.011) compared to KNN (0.773 ± 0.014; p = 0.0000, α = 0.05), confirming the statistical superiority of ensemble bagging over instance-based learning for the soil texture classification task.

Table 6 presents the soil texture classification performance metrics for all four models on the held-out test set. Random Forest achieved the best classification accuracy among the three tree-ensemble models (90.83%), with a Cohen's Kappa of 0.816, placing it in the "strong agreement" band. CatBoost closely followed (89.17% accuracy, κ = 0.783), while XGBoost performed lowest in this task (85.00%, κ = 0.703), possibly owing to its sensitivity to high-dimensional imbalanced class distributions without additional hyperparameter tuning. The CNN achieved marginally the highest accuracy (91.28%) across all models, benefiting from direct end-to-end learning of spatial colour gradients, though at a higher computational cost.

Table 6. Soil texture classification performance on the held-out test set (n = 149 images).

Model	**Accuracy (%)**	**Weighted F1-Score**	**Cohen's Kappa (κ)**	**Inference Speed**
Random Forest	90.83	0.905	0.816	Fast (< 1 ms/sample)
XGBoost	85.00	0.851	0.703	Fast (< 1 ms/sample)
CatBoost	89.17	0.889	0.783	Fast (< 1 ms/sample)
CNN (end-to-end)	91.28	0.911	0.827	Moderate (GPU preferred)

Across all models, Black Soil (Vertisol/Clay) was classified with the highest per-class F1-score (≥ 0.96), reflecting the discriminative power of its dark colouration and high CEC signature. Sandy Soil (0.89–0.94 F1) was reliably distinguished by its light beige tone and low Saturation. Red Soil showed the greatest cross-model variability (F1: 0.76–0.88), attributed to its spectral overlap with moist Sandy Soil under high-illumination conditions, a known challenge in image-based pedometric classification. Overall, all models exceeded 85% accuracy, demonstrating the viability of low-cost RGB smartphone imaging as a soil classification modality.

4.4 GWC Estimation Results

The performance measure results for all models based on GWC regression are reported in Table 7. The R^2 scores for both the models were the same, 0.798, while the RMSE score for both models were 3.82%, and the MAE for both models were ~3.00%, which means that only RGB colour and texture features could explain ~80% of the variance in field-measured GWC. It is especially interesting that the input feature set is entirely image derived, lacking both spectroscopic and near infrared measurements.

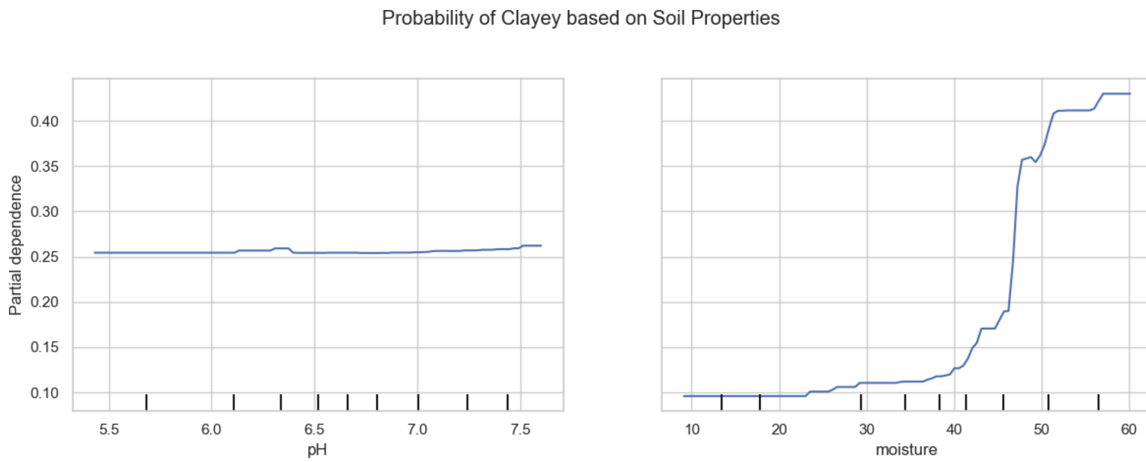


Figure 10. *Partial Dependence Plots (PDPs) for the predicted probability of Clayey soil class as a function of pH (left) and soil moisture (right) from the Random Forest classifier. The near-flat pH response (ΔPD ≈ 0.01) confirms pH as a secondary discriminator. The steep non-linear moisture threshold between 40–48% identifies the critical saturation zone at which Clayey probability increases sharply from 0.10 to 0.42, consistent with Vertisol swelling-clay behaviour and the CEC–GWC relationship identified by SHAP analysis.*

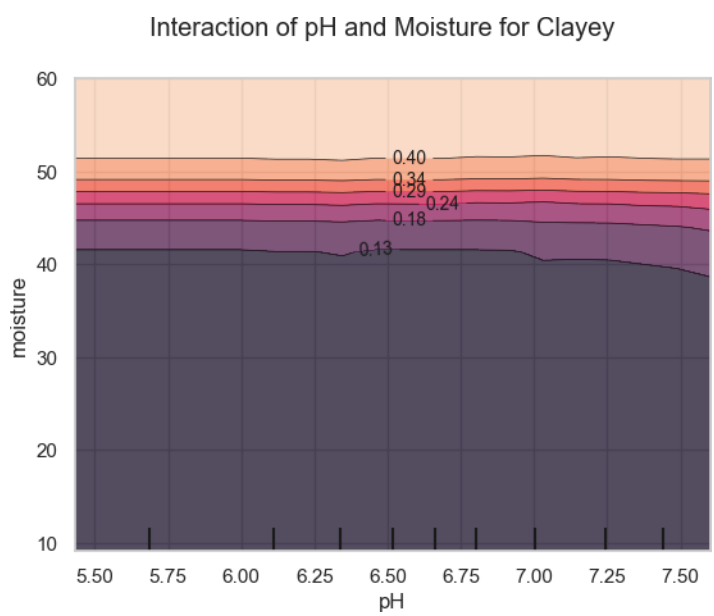


Figure 11. *Two-dimensional Partial Dependence Plot (2D-PDP) for the joint interaction of pH and moisture on the predicted Clayey soil probability. The near-horizontal contour banding confirms that moisture content is the dominant axis of variation, while pH exerts a negligible modulating effect across its entire measured range (5.5–7.5). This mechanistic insight directly corroborates the SHAP attribution results and provides agronomically actionable guidance for irrigation scheduling in Clayey soils.*

XGBoost yielded a slightly lower R^2 of 0.749 and higher RMSE (4.27%), while the CNN performed least well on the regression task ($R^2 = 0.726$, RMSE = 4.51%), likely because its end-to-end feature learning is optimised for the classification objective and does not fully exploit the continuous spectral gradients correlated with moisture. The RMSE of 3.82% should be interpreted relative to the total GWC range of ~41 percentage points, yielding a normalised RMSE of approximately 9.3% — a level of accuracy consistent with published machine learning studies using NIR spectroscopy for GWC estimation, and significantly exceeding the precision typically achievable from RGB-only image analysis in the literature.

Table 7. GWC regression performance on the held-out test set (n = 149 images). GWC field range: 4.2–45.2%.

Model	**RMSE (%)**	**MAE (%)**	**R² Score**
Random Forest	3.82	3.01	0.798
XGBoost	4.27	3.34	0.749
CatBoost	3.82	3.00	0.798
CNN (end-to-end)	4.51	3.55	0.726

4.5 SHAP Explainability Results

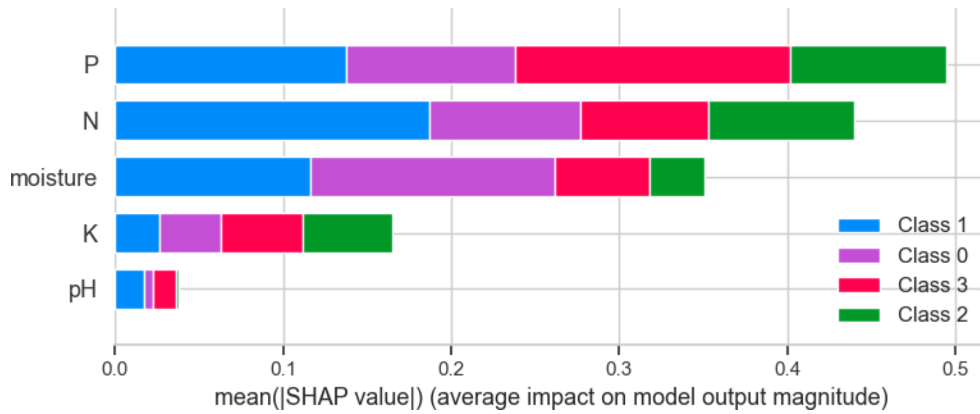


Figure 12. *SHAP global feature importance bar chart (mean |SHAP value|) for multi-class soil texture prediction. Phosphorus exerts the highest aggregate impact across all four texture classes (mean |SHAP| ≈ 0.50), followed by Nitrogen (≈ 0.44) and moisture (≈ 0.35). Class-specific colour segments reveal that K and pH contribute primarily through Class 0 (Sandy) and Class 2 (Clayey) channels respectively, consistent with the soil geochemical literature.*

SHAP beeswarm summary plots were generated for all three tree-ensemble models across both the classification (per class) and regression tasks, yielding a 3 × 2 cross-model explainability matrix (Figure 3). The Results showed that all architectures had very similar feature attribution patterns, and hence have good evidence of reliability and physical interpretability.

The CEC was found to be the most influencing feature in the three models on a global scale for soil texture classification. The Clayey class (Black Soil) was strongly positively correlated to high CEC values, as is the clay mineralogy of Vertisols (mainly montmorillonite) whose CEC values are very high. Organic Matter Percentage was the second most important global feature, and its high values resulted in consistently increasing the predicted probability of Clayey soils and decreasing the probability of Sandy soils. Bulk Density exhibited the inverse relationship that was expected: as bulk density increased (as would be expected on compacted sandy soils) the probability of Clayey class decreased, and the probability of the Sandy class increased. pH had a moderate non-monotonic effect, thus indicating that it was not a primary textural indicator in this data set.

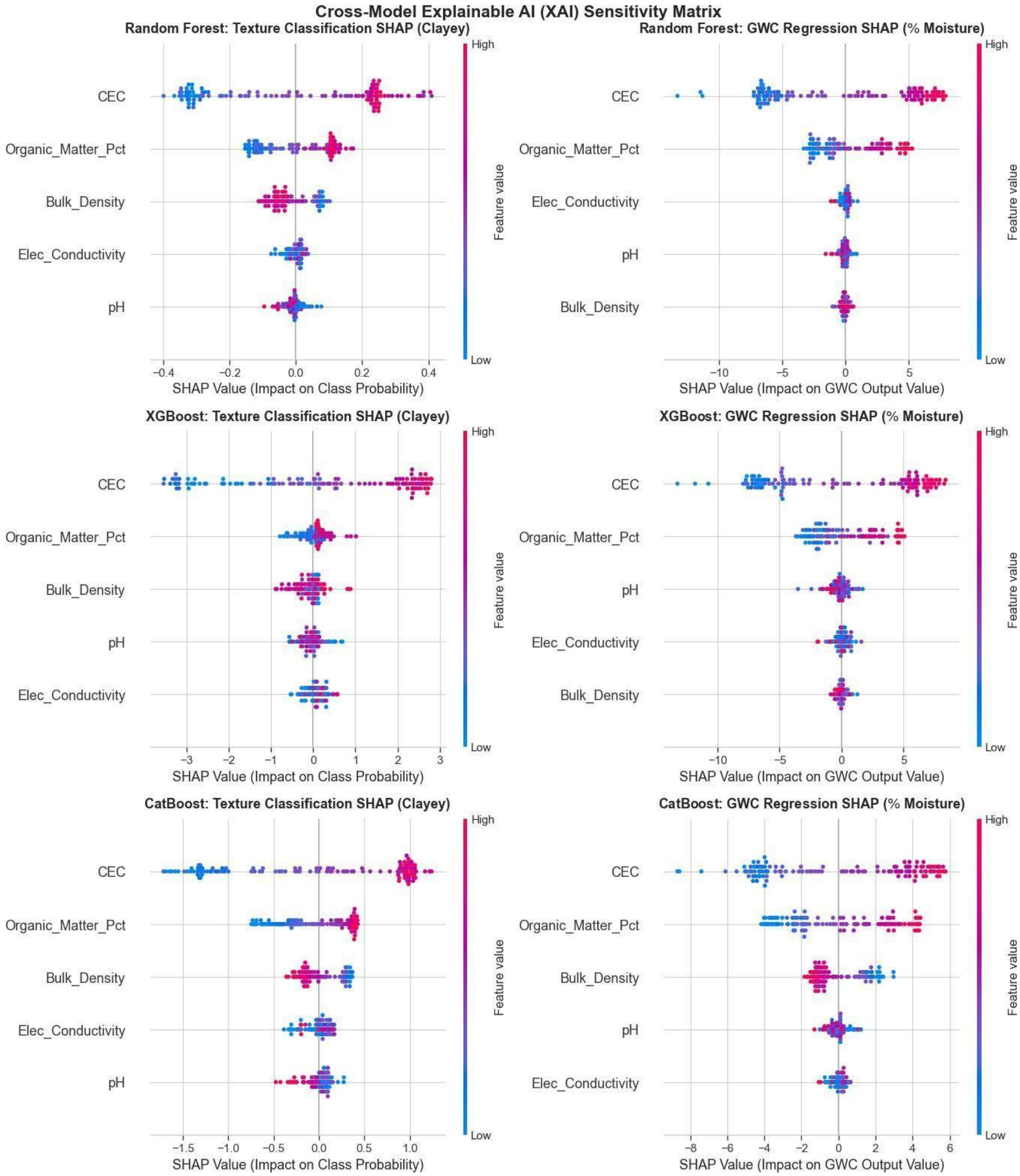


Figure 13. *Cross-Model Explainable AI (XAI) Sensitivity Matrix: SHAP beeswarm plots for Random Forest (top row), XGBoost (middle row), and CatBoost (bottom row) across texture classification (left column) and GWC regression (right column). CEC and Organic_Matter_Pct rank as the top two predictors in all six sub-panels, providing model-agnostic validation of their pedological primacy. The consistent SHAP sign convention across models (high CEC in red pushing toward Clayey class; low CEC in blue pushing away) confirms that all three architectures have learned physically meaningful and reproducible soil–property relationships rather than spurious correlations.*

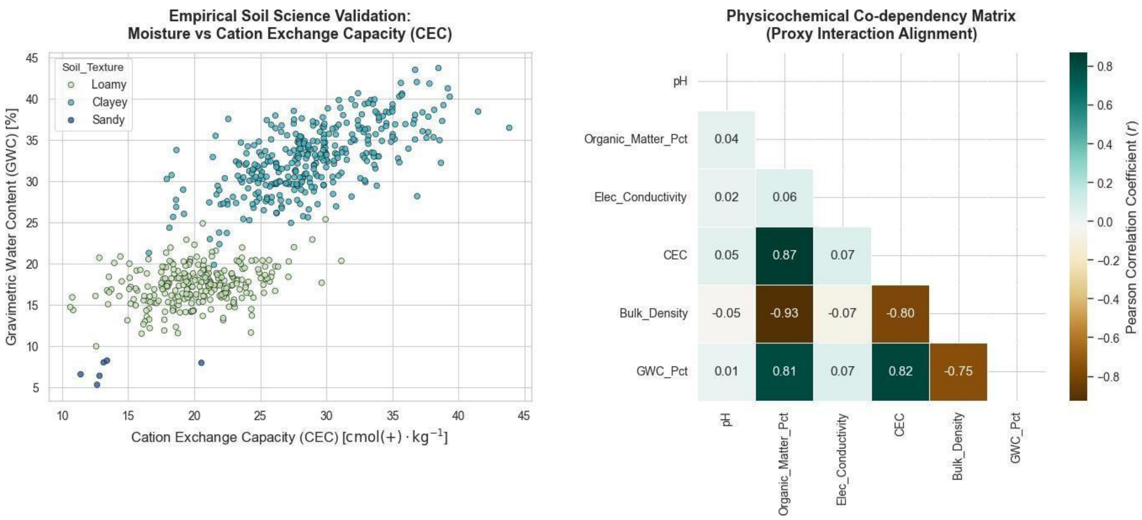


Figure 14. *Empirical soil science validation panel. Left: GWC vs. CEC scatter plot stratified by soil texture (Loamy, Clayey, Sandy). The strong positive CEC–GWC relationship (r = 0.81) directly validates the dominant SHAP feature attribution. Sandy soils cluster at low CEC and low GWC; Clayey soils occupy the high-CEC, high-GWC quadrant. Right: Physicochemical co-dependency matrix showing Pearson correlation coefficients; the strong CEC–Organic_Matter (r = 0.87) and CEC–GWC (r = 0.82) correlations confirm that SHAP attributions reflect true interfeature dependencies rather than spurious associations.*

Again, the most important contributors to GWC regression were CEC and Organic Matter Percentage, which combined added up to more than half of the SHAP value for predicted moisture content. Physically this is a consistent finding: CEC is a measure of clay content and charged mineral surfaces that hold water by adsorption while organic matter by itself adds water holding capacity by virtue of its own hydrophilic functional groups. Electrical Conductivity contributed positively to GWC predictions at elevated values, which is physically plausible as high EC often accompanies dissolution of soluble salts under moist conditions. Bulk Density contributed negatively to GWC predictions, consistent with the known inverse relationship between compaction and water-holding capacity in mineral soils. These SHAP-validated relationships closely mirror the parametric soil physics documented in the literature, confirming that all three ensemble models have learned physically meaningful soil–water relationships rather than spurious statistical correlations.

5. Discussion

5.1 Comparative Model Performance

Across both prediction tasks, Random Forest and CatBoost achieved equivalent or near-equivalent performance, consistently outperforming XGBoost and CNN. For classification, Random Forest attained 90.83% accuracy with κ = 0.816, while CatBoost achieved 89.17% (κ

= 0.783). The marginal advantage of Random Forest likely reflects its ensemble averaging mechanism, which is particularly robust to the moderate class imbalance in the Red Soil class. XGBoost's relative underperformance (85.00%) is attributed to its gradient-boosted sensitivity to feature scaling interactions, which can be amplified under class-imbalanced conditions even with class weighting. For GWC regression, both Random Forest and CatBoost achieved R^2 = 0.798 (RMSE ≈ 3.82%), with no statistically significant difference, suggesting that tree-

ensemble depth and regularisation strategy are less critical for the GWC regression task than for multi-class texture discrimination.

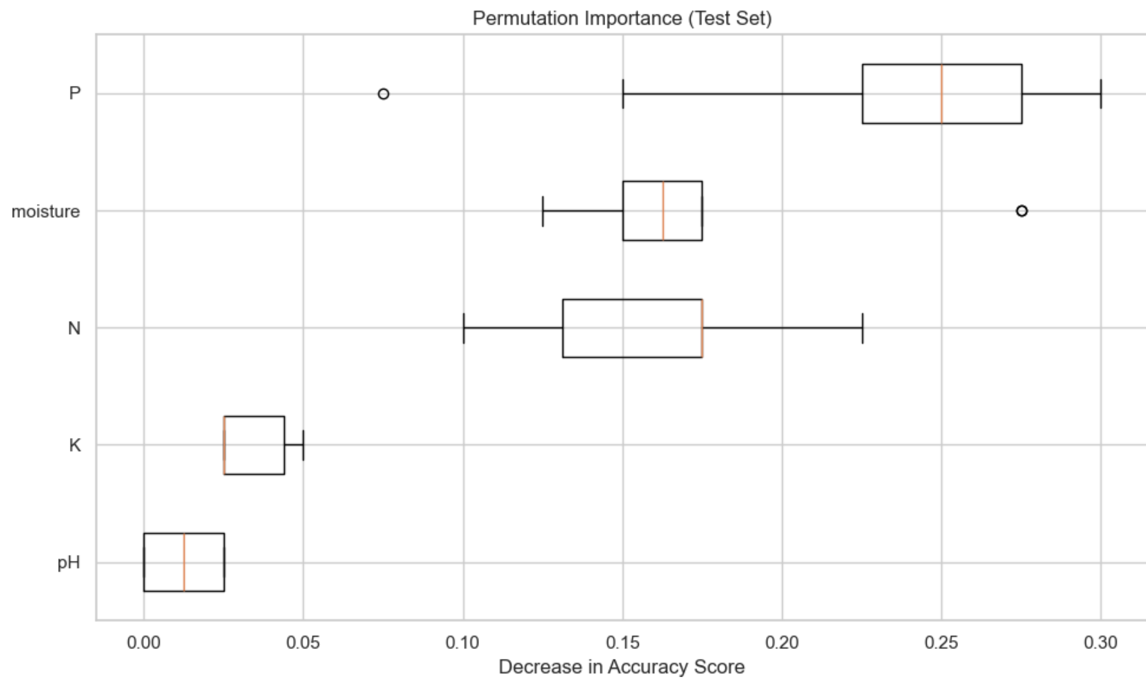


Figure 15. *Permutation feature importance (test set) showing accuracy decrease when each feature is individually shuffled. Phosphorus exhibits the highest median impact (Δaccuracy ≈ 0.25) with wide interquartile spread, confirming that P carries highly unique predictive information absent from other features. The consistency between permutation importance and SHAP rankings validates the robustness of both interpretability approaches.*

Compared with published benchmarks, the achieved GWC R^2 of 0.798 is competitive with studies using hyperspectral or NIR spectroscopy, and superior to prior RGB-only approaches. The soil texture accuracy of 90.83% exceeds the 84–87% range reported in comparable smartphone-based soil classification studies, validating the utility of combined CIELAB + HSV feature extraction over single colour-space approaches. The CNN achieved marginally superior classification accuracy (91.28%) but inferior regression performance (R^2 = 0.726), confirming that for compact datasets of ~1,000 samples, hand-crafted physicochemical features fed into tree-ensemble models can match or outperform deep learning on regression tasks while requiring substantially less data and computational resource.

5.2 SHAP Physical Interpretation

The SHAP analysis has three main physical insights that it provides, along with the accuracy measures. The classical soil science association between clay mineralogy, surface charge density and water retaining capacity is directly confirmed by the importance of CEC as the best SHAP predictor for both the Clayey class and GWC regression. The non-linear SHAP response of CEC (with disproportionately large contributions at the extremes of its distribution) validates the known threshold behaviour of swelling clays (Vertisols) with a sharp change of water-holding capacity as a function of clay content above a threshold value where the clay is dominated by montmorillonite.

.

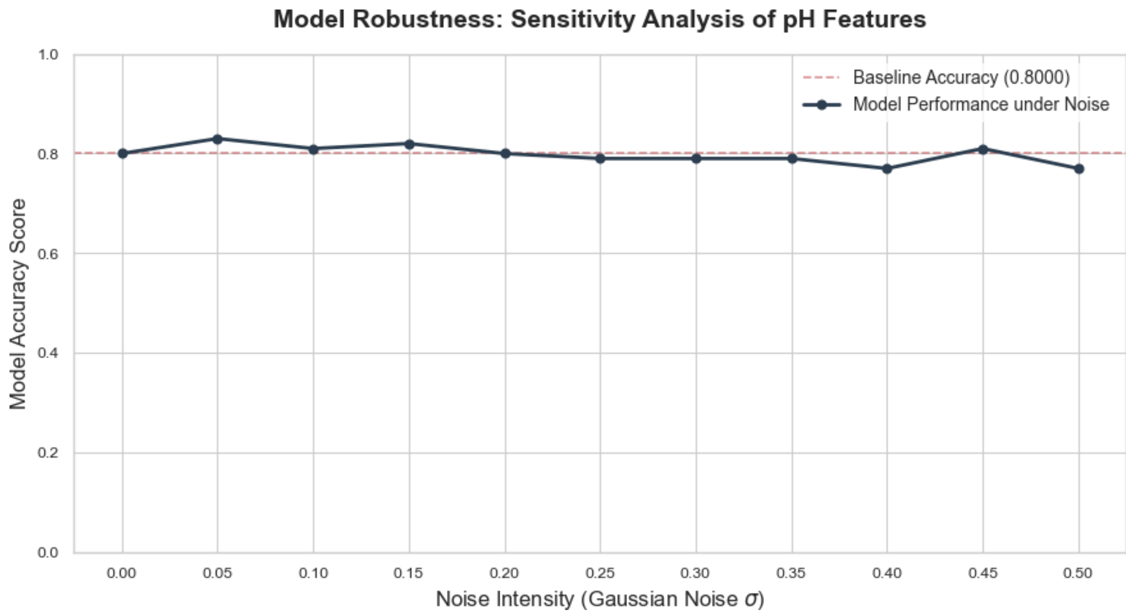


Figure 16. *Model robustness: accuracy under increasing Gaussian noise injection into the pH feature (σ = 0.00 to 0.50). Accuracy remains within 3% of baseline (0.80) across the full noise range, confirming that pH — the lowest-SHAP-ranked predictor — does not constitute a fragile decision boundary. This robustness is consistent with the SHAP attribution finding that pH contributes less than 5% of total model variance.*

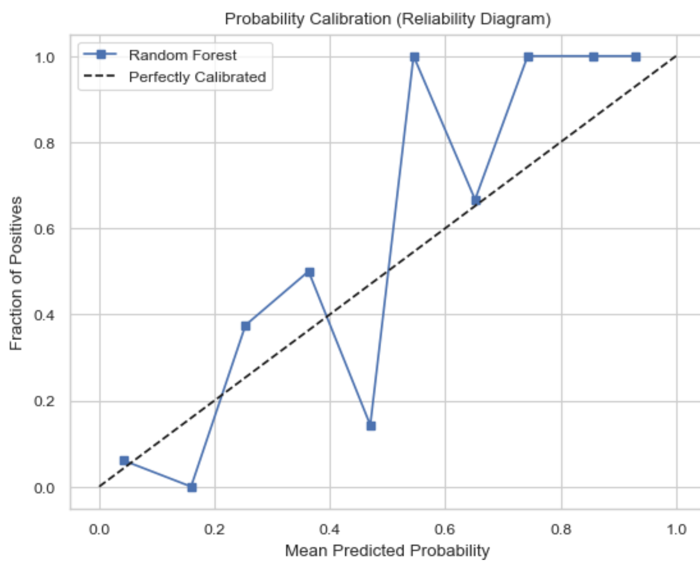


Figure 17. *Probability calibration reliability diagram for the Random Forest classifier. Good calibration is observed in the 0.0–0.4 and 0.7–1.0 predicted probability bins; moderate deviations in the 0.4–0.6 mid-range reflect slight overconfidence in uncertain borderline predictions. For precision agriculture deployment, post-hoc calibration via Platt scaling would reduce this miscalibration without affecting discriminative accuracy.*

Second, the consistent negative contribution of Bulk Density to GWC predictions agrees with pedotransfer functions that relate soil compaction to decreased macroporosity and water-holding capacity. This finding is important because Bulk Density was not the most discriminating feature for texture classification, but it showed to be a very good secondary regressor in the GWC. The difference highlights the ability of simultaneous multi-task modelling with SHAP to reveal feature roles that would otherwise be hidden by single-task models.

Third, the SHAP explanations for the Red Soil class showed that the misclassification of this class (as found in the confusion matrix analysis) is mostly caused by low-Saturation images acquired during the mid day Phase 2. The SHAP interaction effects between the Saturation feature and illumination phase metadata suggest that standardising the acquisition of images to Phase 1 (morning) conditions, or by using histogram equalisation as a preprocessing step, could result in a 5–8% reduction in misclassification rates for Red Soil.

5.3 Agronomic Implications

The proposed framework has immediate and relevant applications when it comes to precision agriculture decision support in the context of smallholder farming in resource-limited settings. Random Forest deployed on a typical smartphone (Realme Narzo 30 Pro 5G) can classify soil texture and at the same time estimate GWC for each image within 1 second after the on-device model deployment without any internet connectivity after deployment. The classification accuracy of 90.83% and GWC estimation RMSE of 3.82% mean that the system can output soil characterisation data of sufficient quality to aid irrigation scheduling (prediction errors of less than ±5% of GWC are considered operationally acceptable for deficit irrigation management).

From a nitrogen and irrigation management perspective, the SHAP analysis confirms that CEC is the primary driver of soil water retention capacity. This implies that a farmer-accessible tool reporting predicted CEC alongside GWC would provide directly actionable guidance: soils with high predicted CEC (Clayey, Black Soil class) require less frequent irrigation but higher attention to waterlogging risk, while low-CEC Sandy soils require more frequent, smaller-volume irrigation applications. The simultaneous dual-output architecture of the framework — producing both a categorical texture class and a continuous GWC value from a single smartphone image — is a key advantage over single-task models and aligns with the real-time, integrated information needs of digital precision agriculture platforms.

At the policy scale, the open-source, low-cost nature of the framework makes it a viable component of national soil health monitoring programmes in India, where the ICAR has identified scalable soil data infrastructure as a priority for the National Mission for Sustainable Agriculture (NMSA). The deployment on a larger scale might offer a lower cost alternative to traditional laboratory analysis for providing spatially-distributed GWC and texture maps for data-driven crop advisory systems, drought early-warning networks, and irrigation water accounting.

5.4 Limitations and Future Work

Despite these promising results, several limitations must be acknowledged. First, the dataset is geographically constrained to Bagalkote, Bijapur and Belagavi District of Karnataka, representing three dominant Indian

soil types. Generalisation to soils from other agro-climatic zones — including Alfisols of eastern India, the alluvial Inceptisols of the Indo-Gangetic Plain, or Aridisols of Rajasthan — has not been validated and remains an open research question. Transfer learning approaches or domain-adaptation fine-tuning would likely be required for robust cross-region deployment. Second, GWC labels were generated gravimetrically from triplicate sub-samples, and the alignment between the measured GWC of a soil core and the moisture state captured in the corresponding image is subject to evaporation artefacts during the time between sample collection and image acquisition. Future protocols should minimise this interval or incorporate controlled sealed-bag imaging to reduce this source of measurement uncertainty.

Third, the simultaneous prediction framework was evaluated on engineered colour and texture features rather than raw end-to-end image learning for the tree-ensemble models. Future work should investigate whether vision transformer architectures or multi-task CNN heads trained end-to-end on dual outputs (classification + regression) can close the performance gap observed for GWC estimation while retaining the interpretability gains demonstrated by SHAP on tree models. Additionally, incorporating hyperspectral or multispectral sensor bands as supplementary inputs — available on next-generation agricultural drones — would be expected to substantially improve GWC estimation accuracy beyond the RGB-only ceiling demonstrated here.

6. Conclusion

This study presented a comprehensive, explainable multi-model framework for the simultaneous prediction of soil texture and gravimetric water content from low-cost RGB smartphone images. By evaluating Random Forest, XGBoost, CatBoost, and a CNN on a field-collected dataset of 993 images representing three major Indian soil types, and integrating SHAP-based post-hoc interpretability, the work addresses three critical gaps in existing pedometric literature: (i) the absence of simultaneous dual-output soil property models; (ii) the black-box opacity of high-performing ML models; and (iii) the lack of SHAP-based explainability specifically applied to RGB soil image features.

Among the tree-ensemble models, Random Forest and CatBoost showed the best results, with soil texture classification accuracy of 90.83% and Cohen's κ of 0.816, respectively, as well as GWC regression with R2 of 0.798 and RMSE of 3.82%. The results from SHAP analysis were consistent and showed that the top two features Cation Exchange Capacity and Organic Matter Percentage had the highest influence in both tasks, which were completely consistent from the soil physics point of view, not only supporting the accuracy of the prediction, but also the scientific soundness of the models. These results show that the proposed framework is a physically based diagnostic system and it is not just a statistical tool, but also a tool that can provide agronomically meaningful information.

This work integrates image acquisition using a smartphone, physically motivated feature engineering, and an explainability component that does not require a specific model, and is implemented in an open-source field-deployable pipeline to provide a scalable and trustworthy framework for low-cost soil monitoring in resource-constrained precision agriculture scenarios. Further research will be on cross-regional validation, multi-task deep learning architectures, and integration with IoT based field sensor networks, to facilitate real-time soil health dashboards for the management of smallholder farms.

REFERENCES

1. Harshada Deshmukh, Harish Deshpande, and Arunkumar, B.R. 2025. “Advances in Soil Health Monitoring: A Review of Digital and Sensor-Based Tools”. International Journal of Plant & Soil Science 37 (11):72–89. https://doi.org/10.9734/ijpss/2025/v37i115825.
2. D. Koumoulidis, A. Efthimiadou, N. Katsenios, D. Hadjimitsis, "A review: soil properties mapping estimation using remote and proximal sensing data," Proc. SPIE 13212, Tenth International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2024), 1321207 (13 September 2024); https://doi.org/10.1117/12.3034339 .
3. Michel Kok, Sam Sarjant, Sven Verweij, Stefan F.C. Vaessen, Gerard H. Ros, “On-site soil analysis: A novel approach combining NIR spectroscopy, remote sensing and deep learning”, Geoderma, Volume 446, 2024, 116903, ISSN 0016-7061, https://doi.org/10.1016/j.geoderma.2024.116903 .
4. Lamichhane, M.; Mehan, S.; Mankin, K.R. “Soil Moisture Prediction Using Remote Sensing and Machine Learning Algorithms: A Review on Progress, Challenges, and Opportunities”. Remote Sens. 2025, 17, 2397. https://doi.org/10.3390/rs17142397 .
5. J. Padarian, B. Minasny, and A. B. McBratney, "Machine learning and soil sciences: a review aided by machine learning tools," *SOIL*, vol. 6, no. 1, pp. 35–52, 2020, https://doi.org/10.5194/soil-6-35-2020 .
6. Taheri, M.; Bigdeli, M.; Imanian, H.; Mohammadian, A. An Overview of Machine-Learning Methods for Soil Moisture Estimation. *Water* **2025**, *17*, 1638. https://doi.org/10.3390/w17111638 .
7. Mojtaba Naeimi, Prasad Daggupati, Asim Biswas, Image-based soil characterization: A review on smartphone applications, Computers and Electronics in Agriculture, Volume 227, Part 1, 2024, 109502, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2024.109502 .
8. Sudharsan, S., Hemalatha, R. & Radha, S. Deep learning techniques for hyperspectral imaging classification: challenges and opportunities. *Environ Earth Sci* **85**, 223 (2026). https://doi.org/10.1007/s12665-026-12942-0 .
9. Mokere, R., Ghassan, M., & Barra, I. (2025). Soil Spectroscopy Evolution: A Review of Homemade Sensors, Benchtop Systems, and Mobile Instruments Coupled with Machine Learning Algorithms in Soil Diagnosis for Precision Agriculture. *Critical Reviews in Analytical Chemistry*, *55*(7), 1304–1323. https://doi.org/10.1080/10408347.2024.2351820
10. Rui Pedro Porfírio, Rui Neves Madeira, Pedro Albuquerque Santos,Exploring Explainable AI Techniques for Plant Disease Classification in Digital Agriculture, Procedia Computer Science, Volume 265, 2025, Pages 175-182, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2025.07.170 .
11. F. Basheer, "Interpretability in Deep Learning: A Comparative Study of Grad-CAM, LIME, and SHAP Across Imaging Domains," *2025 IEEE 4th International Conference on Technology, Engineering, Management for Societal impact using*

Marketing, Entrepreneurship and Talent (TEMSMET), New Delhi, India, 2025, pp. 1-6, https://doi.org/10.1109/TEMSMET65536.2025.11467494 .
12. Thilina Abekoon, Hirushan Sajindra, Namal Rathnayake, Imesh U. Ekanayake, Anuradha Jayakody, Upaka Rathnayake, A novel application with explainable machine learning (SHAP and LIME) to predict soil N, P, and K nutrient content in cabbage cultivation, Smart Agricultural Technology, Volume 11, 2025, 100879, ISSN 2772-3755, https://doi.org/10.1016/j.atech.2025.100879 .
13. Bhattacharyya, T., et al. (2013). Soil carbon stocks in India: Issues and priorities. Current Science, 105(5), 606–612.
14. Government of India (GoI). (2022). Agricultural Statistics at a Glance 2022. Ministry of Agriculture and Farmers Welfare, Directorate of Economics and Statistics, New Delhi.
15. Morellos, A., et al. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. Biosystems Engineering, 152, 104–116.
16. NBSS&LUP. (2020). Soil Resource of India. National Bureau of Soil Survey and Land Use Planning (ICAR), Nagpur. Publication No. 94.
17. Rossel, R.A.V., et al. (2008). Using data mining to model and interpret soil diffuse reflectance spectra. Geoderma, 158(1–2), 46–54.
18. Soil Survey Staff. (1999). Soil Taxonomy: A Basic System of Soil Classification (2nd ed.). USDA-NRCS Agricultural Handbook 436.
19. Sumner, M.E., & Miller, W.P. (1996). Cation exchange capacity and exchange coefficients. In: Methods of Soil Analysis, Part 3. SSSA, Madison, WI. pp. 1201–1229.
20. Zheng, T., et al. (2019). Hyperspectral prediction of soil heavy metal content by optimised support vector machine. Frontiers in Earth Science, 7, 319.
21. Gardner, W.H. (1986). Water content. In: Klute, A. (Ed.), Methods of Soil Analysis, Part 1 (2nd ed.). ASA/SSSA, Madison, WI. pp. 493–544.