Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

Machine Learning-Based Soil Productivity Estimation from Multitemporal Spectral Features

Manju Kumari[1], Madhu Nimesh[2*]
[1]*Assistant Professor, Department of Electronics Engineering J C Bose University of Science and Technology, YMCA, Faridabad -121006, Haryana (INDIA)*
[2*]*Independent Researcher*

ARTICLEINFO

Received: 03 Nov 2024

Revised: 17 Dec 2024

Accepted: 29 Dec 2024

ABSTRACT

Reliable spatial information on soil condition is important for precision agriculture in Haryana, where intensive cultivation and spatial variation in soil fertility create a need for more detailed assessment than can be obtained from sparse point-based testing alone. This study develops a multitemporal Sentinel-2 machine-learning framework for mapping soil productivity across agricultural areas of Haryana, India. The framework integrates cloud- and shadow-screened surface reflectance, bare-soil identification, temporal compositing, spectral-feature engineering, feature optimization, multi-property regression, and an agronomically interpreted Soil Productivity Index (SPI). Soil organic carbon (SOC), nitrogen (N), phosphorus (P), potassium (K), and pH are treated as complementary indicators of soil fertility and chemical suitability. Four regression approaches partial least squares regression (PLSR), support vector regression (SVR), random forest (RF), and extreme gradient boosting (XGBoost)—are compared under spatially independent validation. Feature selection is performed within the model-development process to avoid information leakage, and the final predictions are evaluated using R^2, RMSE, MAE, and RPD. The predicted indicators are transformed into standardized suitability scores and integrated into an SPI using agronomically justified response functions, with sensitivity analysis of alternative weighting schemes. SHAP analysis is used to interpret the contribution of image-derived predictors, while resampling-based uncertainty estimation is used to identify areas in which the productivity estimates are less reliable. The study is positioned as an integrated framework rather than as an algorithmic novelty claim: Sentinel-2, machine learning, feature selection, SHAP, and composite soil indices are established individually, whereas their coordinated use for spatially validated, multi-indicator productivity assessment in Haryana provides the principal methodological contribution. The final manuscript reports empirical performance only from the Haryana ground-reference dataset and independent spatial validation actually used in the experiment.

Keywords: Haryana; soil productivity; Sentinel-2; digital soil mapping; soil organic carbon; nitrogen; phosphorus; potassium; pH; machine learning; SHAP; uncertainty; precision agriculture.

I. INTRODUCTION

Soil is a fundamental component of agricultural production and plays a central role in nutrient cycling, water regulation, carbon storage, and long-term ecosystem sustainability. The productivity of agricultural soils is governed by the interaction of several physical, chemical, and biological properties rather than by a single soil attribute. Soil organic carbon (SOC), nitrogen (N), phosphorus (P), potassium (K), pH, electrical conductivity, texture, and related properties collectively influence nutrient availability, root development, water retention, and the capacity of soil to support crop growth. Conventional determination of these properties relies primarily on laboratory-based sampling and chemical analysis, which provides reliable measurements but is relatively expensive, labor intensive, and spatially sparse. Consequently, there is an increasing requirement for rapid, non-destructive, spatially continuous approaches capable of characterizing soil conditions over agricultural landscapes.

Spectral sensing has emerged as a promising approach for overcoming several limitations associated with conventional soil assessment. Visible near-infrared (Vis–NIR), near-infrared (NIR), and mid-infrared (MIR) spectroscopy can capture soil-specific spectral responses associated with organic and mineral constituents and can therefore be coupled with chemometric and machine-learning techniques for quantitative estimation of soil properties. Barra et al. [1] reviewed the development of soil spectroscopy and demonstrated that the integration of spectral preprocessing, variable selection, chemometrics, and machine learning can provide rapid and cost-effective estimation of several soil physical and chemical properties. However, laboratory and proximal spectroscopy generally provide measurements at discrete locations and require dedicated acquisition systems. The transition from proximal spectroscopy to airborne and satellite remote sensing therefore provides an important opportunity for extending spectral soil assessment to larger agricultural areas.

Recent advances in Earth-observation platforms, particularly Sentinel-2 [16], have strengthened the feasibility of spatial soil-property estimation. Sentinel-2 provides multispectral observations covering the visible, near-infrared, red-edge, and shortwave-infrared regions, which contain information relevant to soil organic matter, mineral composition, moisture and related surface characteristics. Wang et al. [2] evaluated a large continental soil spectral library containing 37,540 spectra across 350–2500 nm and compared several machine-learning algorithms and spectral preprocessing approaches for SOC estimation. Their results demonstrated the importance of spectral preprocessing, nonlinear learning approaches, and shortwave-infrared information for soil carbon prediction. Nevertheless, the study also highlighted the influence of soil moisture, vegetation and plant residues on spectral prediction, indicating that the representation and selection of spectral observations are critical for reliable soil-property modelling [2]. These findings suggest that simply applying machine-learning algorithms to all available image bands may not provide an optimal solution and that systematic image preprocessing and spectral feature optimization are required.

The availability of harmonized soil observations provides an important basis for developing and evaluating such image-based approaches. The LUCAS 2015 TOPSOIL database contains 21,859 geographically referenced observations across the European Union and includes measurements of organic carbon, total nitrogen, phosphorus, extractable potassium, pH, electrical conductivity, texture, and multispectral reflectance [3]. The large spatial coverage and combination of laboratory soil measurements and spectral information make LUCAS particularly valuable for developing reproducible machine-learning models for remote soil assessment. Recent research has demonstrated the potential of combining LUCAS observations with Sentinel-2 imagery for large-scale SOC mapping. For example, van Wesemael et al. [4] developed a European SOC monitoring system using Sentinel-2 soil-reflectance composites and the LUCAS soil database. Their approach employed multitemporal Sentinel-2 information and provided spatially explicit uncertainty estimates; however, the reported continental-scale performance remained moderate, with R2 values of approximately 0.41 for croplands and 0.28 for permanently vegetated areas [4]. These findings demonstrate both the potential and the limitations of satellite-based soil-property prediction and emphasize the importance of temporal compositing, spatial validation, and uncertainty assessment.

The temporal representation of satellite observations is particularly important for agricultural soils. Soil spectral response can vary substantially with moisture, crop residue, vegetation cover, illumination, and surface management. Consequently, a single Sentinel-2 acquisition may not adequately represent the underlying soil condition. Recent research using multitemporal Sentinel-2 observations has demonstrated the value of bare-soil compositing for SOC prediction and has shown that temporal aggregation can provide more stable representations of soil reflectance than individual image dates. This is consistent with the broader spectral literature, which identifies preprocessing and feature selection as important determinants of soil-property prediction performance [1], [2]. In addition, recent Sentinel-2-based digital soil mapping studies have demonstrated that hybrid models incorporating spatial information can outperform individual machine-learning models, emphasizing the importance of considering spatial autocorrelation and validation design when evaluating remote-sensing soil models [5].

Despite these developments, most existing image-based studies concentrate on the prediction of individual soil properties, particularly SOC, rather than developing an integrated assessment of soil productivity. Although predicting SOC is important, SOC alone cannot adequately represent nutrient availability or chemical constraints affecting crop production. A soil having high SOC, for example, may still have inadequate phosphorus or potassium availability or an unfavorable pH. Consequently, an assessment intended to characterize soil productivity should integrate multiple complementary soil indicators rather than rely on a single predicted variable. This requirement

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

motivates a transition from the conventional formulation “Image to Individual Soil Property” toward “Image to {SOC,N,P,K,pH}to Soil Productivity”.

A related development in the literature is the use of composite Soil Quality Indices (SQIs). The Soil Management Assessment Framework proposed by Andrews et al. [6] established a quantitative approach involving indicator selection, interpretation/scoring, and integration into a composite index. More recently, spectral and hyperspectral sensing has been extended from individual soil-property prediction toward direct soil-quality assessment. Gözükara et al. [7], for example, demonstrated the potential of Vis–NIR and portable X-ray fluorescence spectra for estimating soil properties and a composite soil-quality index. More significantly, Majeed and Das [8] developed a large-scale soil-quality mapping approach using airborne AVIRIS-NG hyperspectral data and constructed SQIs from 16 soil properties measured at 101 locations. Their work demonstrates the feasibility of using remotely sensed spectral information to estimate a composite soil-quality measure at landscape scale. However, such approaches generally rely on hyperspectral or proximal sensing and do not directly address the development of an operational, multispectral satellite-based soil productivity framework integrating multiple productivity-relevant soil indicators.

Furthermore, high predictive accuracy alone does not necessarily guarantee reliable spatial decision support. Spatial dependence between training and test observations can result in overly optimistic estimates when conventional random cross-validation is used. Recent Sentinel-2 soil mapping research has demonstrated the influence of spatial resolution and spatial autocorrelation on SOC prediction performance [5]. Therefore, spatially independent or spatially grouped validation is required to assess whether a model can generalize beyond locations that are geographically similar to its training observations. In addition, complex machine-learning models can provide limited information about why a particular prediction is obtained. Explainable artificial intelligence, particularly SHAP (SHapley Additive exPlanations), provides a theoretically grounded framework for quantifying the contribution of individual predictors to model outputs [9]. Incorporating explainability can therefore help identify the spectral features most strongly associated with predicted soil conditions. Similarly, spatial uncertainty should be reported alongside productivity estimates so that regions with low confidence are not interpreted in the same manner as reliable predictions.

Based on these observations, a significant methodological opportunity remains to integrate multitemporal multispectral image processing, optimized spectral-feature extraction, multi-property machine learning, productivity-oriented index construction, spatially robust validation, explainable modelling, and uncertainty assessment within a single reproducible framework. The novelty of the present study is therefore not claimed from the individual use of Sentinel-2, machine learning, SHAP, or soil-quality indexing, as these techniques have already been established independently [10]-[12]. Rather, the contribution lies in their integration into an image-derived, productivity-oriented framework that transforms spatially continuous multispectral observations into multiple soil-property estimates and subsequently into an interpretable Soil Productivity Index (SPI).

The organization of this manuscript is as, in Section I introduction being given, Section II provides the literature review which gives the yields research gaps, with the targeting these research gaps objectives put forward. Section III gives the methodology part, sections IV has the results and discussion. Finally, the conclusion and future scope given in section V and VI, and the references has been given.

II. LITERATURE REVIEW

A. Spectral and Image-Based Soil Characterization

Conventional assessment of soil properties relies primarily on field sampling followed by laboratory-based chemical and physical analyses. Although such methods provide reliable measurements, their application over large agricultural areas is constrained by sampling density, cost, processing time, and the difficulty of representing spatial heterogeneity. Consequently, spectral sensing has emerged as an important alternative for rapid and spatially distributed soil characterization. Visible–near-infrared (Vis–NIR), near-infrared (NIR), and mid-infrared (MIR) spectroscopy provide information related to organic matter, minerals, moisture, and other soil constituents and can be coupled with chemometric and machine-learning methods for quantitative soil-property estimation.

Barra et al. [1] reviewed the use of soil spectroscopy with preprocessing, chemometrics, machine learning, and variable-selection techniques and concluded that infrared spectroscopy can provide rapid, quantitative, and environmentally favourable soil diagnostics. Importantly, their review demonstrated that the predictive performance of soil spectroscopy depends not only on the learning algorithm but also on spectral preprocessing and variable selection. This is particularly relevant to image-based soil assessment because remotely sensed datasets contain correlated spectral variables and derived indices that can introduce redundancy if they are directly supplied to a model.

Wang et al. [2] provided further evidence of the potential of spectral machine learning using a continental-scale library of 37,540 soil spectra covering 350–2500 nm. Seven machine-learning algorithms and four spectral preprocessing approaches were evaluated for SOC prediction. LSTM produced R2=0.96 for the complete spectral dataset, while first-derivative preprocessing improved several conventional algorithms. The study also demonstrated that soil moisture, vegetation and plant residues affect SOC prediction and identified the importance of the shortwave-infrared region for airborne and spaceborne soil sensing.

These studies establish the fundamental feasibility of spectral soil assessment but also reveal an important limitation: high accuracy obtained using laboratory or hyperspectral measurements cannot automatically be transferred to multispectral satellite observations. Satellite imagery is affected by atmospheric conditions, illumination, soil moisture, vegetation, residue, surface roughness, and spatial mixing. Consequently, an operational image-based soil assessment framework requires systematic preprocessing, selection of representative soil observations, and optimization of image-derived spectral features.

B. Sentinel-2 and Multitemporal Soil Mapping

The Sentinel-2 mission provides multispectral observations in the visible, red-edge, near-infrared, and shortwave-infrared regions with relatively high spatial resolution and frequent revisit capability. These characteristics make Sentinel-2 particularly attractive for agricultural soil monitoring. The LUCAS 2015 TOPSOIL dataset further strengthens this opportunity by providing a large, geographically referenced soil database containing laboratory measurements of organic carbon, nitrogen, phosphorus, potassium, pH, electrical conductivity, texture, and multispectral reflectance [3]. The dataset therefore provides a valuable reference source for developing and validating satellite-based soil-property models.

A particularly important recent contribution is the European SOC monitoring system developed by van Wesemael et al. [4]. Their WorldSoils framework combined Sentinel-2 soil-reflectance composites with the LUCAS soil database to provide continuous SOC prediction across Europe. The system used median Sentinel-2 reflectance composites over a three-year period and produced spatially explicit uncertainty information. At continental scale, the reported SOC prediction performance was R^2=0.41 for croplands and R^2=0.28 for permanently vegetated areas, demonstrating both the feasibility and the limitations of operational satellite SOC monitoring.

Temporal representation is particularly important for agricultural soils because soil surfaces are not continuously exposed. Vegetation, crop residues, soil moisture and management operations can alter the observed spectral response. Zhu et al. [4] investigated this issue explicitly using *12 Sentinel-2 bare-soil observations collected between 2017 and 2022* over a 4,700-km² cropland region in northeast China. The individual-date models exhibited substantial temporal instability, with R^2 values ranging from 0.30 to 0.67. A geometric-median bare-soil composite produced R^2=0.64 and RMSE =2.24 g kg^{-1}, outperforming 11 of the 12 individual-date models. Incorporating SOC-related spectral indices produced a further 6.5% improvement in prediction accuracy.

These findings provide a strong basis for adopting multitemporal bare-soil Sentinel-2 observations in the proposed study. However, Zhu et al. [4] focused primarily on SOC. The broader question of whether such a strategy can be extended to multiple productivity-relevant soil indicators and subsequently transformed into a spatial soil-productivity measure remains insufficiently investigated.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

C. Machine Learning and Spectral Feature Optimization

Machine-learning techniques have become increasingly important for soil-property estimation because the relationships between spectral response and soil properties are generally nonlinear and affected by interactions among organic matter, minerals, texture, moisture, and surface conditions. PLSR provides an established chemometric baseline, whereas SVR, RF, gradient-boosting and neural-network models can represent nonlinear relationships.

The comparative analysis by Wang et al. [2] demonstrated that model performance depends strongly on the spectral representation and preprocessing strategy. Their results showed that even conventional models could benefit substantially from appropriate preprocessing, indicating that increasing algorithmic complexity is not necessarily sufficient to improve prediction [2]. Similarly, the soil-spectroscopy review by Barra et al. [1] emphasized the importance of preprocessing and variable-selection strategies.

This issue becomes particularly relevant when multispectral bands are combined with normalized differences, band ratios and other spectral indices. Such feature engineering can increase the predictor space while simultaneously introducing multicollinearity and redundant information. Therefore, a robust framework should distinguish between *feature generation* and *feature selection*. Variance filtering, correlation analysis, mutual information, recursive feature elimination, PCA, and tree-based importance can be used to identify a compact set of informative predictors.

Another important consideration is validation. Conventional random training–testing splits may produce optimistic estimates when spatially adjacent observations occur in both training and testing subsets. For a spatial soil-productivity application, the relevant question is not simply whether the model predicts randomly withheld samples, but whether it can generalize to geographically distinct locations. Consequently, spatially grouped or spatial-block cross-validation provides a more rigorous assessment of model transferability.

D. From Individual Soil Properties to Soil Quality

Most remote-sensing soil studies have concentrated on individual properties, particularly SOC. Although SOC is an important indicator of soil health and functioning, it cannot independently represent the nutrient and chemical conditions affecting agricultural productivity. Nitrogen, phosphorus and potassium directly influence plant nutrition, while pH affects nutrient availability and microbial activity. Thus, a soil-productivity assessment should ideally integrate several complementary soil indicators.

The development of composite Soil Quality Indices (SQIs) provides an established basis for such integration. Andrews, Karlen and Cambardella [6] proposed the Soil Management Assessment Framework (SMAF), which provides a systematic approach for selecting soil indicators, converting them into scores, and integrating them into an overall soil-quality assessment. This conceptual structure is particularly relevant to the proposed research because it separates ***measurement*, *indicator scoring*, and *index construction*,** allowing remotely predicted soil properties to be transformed into an interpretable composite productivity measure.

Spectral approaches have subsequently been extended from individual properties to SQI prediction. Gözükara et al. [7] investigated soil-property and SQI prediction using Vis–NIR and portable X-ray fluorescence spectra. Their results demonstrated that combined spectral information could improve prediction relative to individual sensing sources and that a Cubist model could be used for SQI estimation.

An especially relevant contribution is the work of Majeed and Das [8], who developed a large-scale soil-quality mapping approach using airborne AVIRIS-NG hyperspectral data. Their SQIs were constructed from 16 soil properties measured at tentatively around 101 locations, and chemometric models were used to estimate SQI from laboratory and airborne hyperspectral reflectance. This study provides strong evidence that a composite soil-quality measure can be inferred from remotely sensed spectral information.

However, soil quality and soil productivity should not be treated as identical concepts. SQI is generally concerned with the capacity of soil to perform defined soil functions, whereas productivity is more directly related to the capacity of soil to support plant growth under agricultural conditions. Therefore, the proposed study extends the concept from

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

spectral SQI toward an agronomically interpreted Soil Productivity Index (SPI) based on multiple remotely predicted soil indicators.

E. Explainability and Uncertainty in Soil Productivity Mapping

High predictive accuracy alone is insufficient when machine-learning models are intended to support spatial agricultural decisions. Complex models may provide accurate predictions without explaining which image features control those predictions. Explainable artificial intelligence can address this limitation by providing feature-level explanations.

Lundberg and Lee [9] proposed SHAP (SHapley Additive exPlanations), a unified framework for interpreting model predictions by assigning contribution values to individual features. SHAP is therefore appropriate for identifying which Sentinel-2 bands, spectral ratios, or derived indices contribute most strongly to predictions of SOC, N, P, K, and pH.

Uncertainty is a second important consideration. The European Sentinel-2/LUCAS SOC monitoring framework of van Wesemael et al. [3] explicitly provided pixel-level uncertainty information, illustrating that prediction confidence varies spatially even when a continuous soil map is available. Therefore, the proposed framework treats uncertainty as an integral part of the final productivity product rather than as an optional statistical analysis.

F. Critical Literature Survey

The principal studies relevant to the proposed research are summarized in Table I. The table combines the earlier literature-survey framework with the strongest papers identified for the present research direction.

TABLE I : Critical Literature Survey On Spectral/Image-Based Soil Assessment

Author (s) & Year	Title / Topic	Objectives / Research Purpose	Methodology / Approach	Key Findings	Limitations / Research Gaps
Barra et al., 2021 [1]	Soil spectroscopy, chemometrics and ML	Review spectral methods for soil diagnosis	NIR/Vis-NIR/MIR, preprocessing, chemometrics, ML	Spectroscopy can provide rapid quantitative soil diagnosis; preprocessing and variable selection are important	Review-level study; no unified satellite productivity framework
Wang et al., 2022 [2]	Hyperspectral reflectance and ML for SOC	Assess potential of laboratory spectra for airborne/spaceborne SOC sensing	7 ML algorithms, 4 preprocessing methods, radiative-transfer simulation	LSTM achieved R2=0.96; preprocessing affected conventional models; SWIR important	Primarily SOC; laboratory spectra and simulated satellite observations; limited multi-property productivity assessment
Gözükara et al., 2022 [7]	Vis-NIR and pXRF-based SQI	Predict soil properties and SQI	Vis-NIR + pXRF + Cubist	Combined spectra improved prediction; Cubist could predict SQI	Proximal/profile-scale sensing; not multitemporal satellite imagery
Majeed & Das, 2024 [8]	Airborne hyperspe	Map SQI across	AVIRIS-NG + 16 soil properties +	Demonstrated landscape-scale SQI estimation	Hyperspectral airborne data; does not address Sentinel-

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

	ctral SQI mapping	different land uses	chemometric models	from hyperspectral data	2 multispectral productivity mapping
Zhu et al., 2024 [4]	Single-date vs multitemporal Sentinel-2 SOC	Evaluate temporal compositing strategies	12 bare-soil images; PLSR, Cubist, RF; 100 bootstrap validations	Single-date R2=0.30–0.67; geometric-median composite R2=0.64, RMSE 2.24 g/kg; indices improved accuracy 6.5%	Primarily SOC; future extension to integrated soil indicators suggested
wang et al., 2024 [3]	European Sentinel-2/LUCAS SOC monitoring	Develop continental SOC monitoring system	Sentinel-2 temporal composites + LUCAS + spectral/DSM models	Cropland R2=0.41; uncertainty provided at pixel level	SOC only; continental performance remains moderate; no productivity index
Proposed study	Image-Based Soil Productivity Assessment Using Multitemporal Sentinel-2 Imagery and Machine Learning	Integrate soil-property prediction into productivity assessment	Multitemporal Sentinel-2 + feature optimization + PLSR/SVR/RF/XGBoost + SPI + SHAP + uncertainty + spatial validation	To be experimentally established	Requires rigorous empirical validation

G. Research Gaps Identified

The critical review identifies the following gaps.

Gap	Identified limitation
Gap 1 – Individual soil-property prediction	Most studies focus on one soil property, particularly SOC, rather than several properties together.
Gap 2 – Limited productivity-oriented integration	Existing Soil Quality Index (SQI) approaches do not necessarily represent the actual productivity potential of agricultural soil.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

Gap 3 – Temporal instability	Soil-property predictions from a single Sentinel-2 image can vary with time and surface conditions.
Gap 4 – Spectral feature redundancy	Sentinel-2 provides many related spectral features, which may contain redundant information.
Gap 5 – Limited spatial validation	Random data splitting can give overly optimistic results when nearby samples are spatially related.
Gap 6 – Limited interpretability	Complex machine-learning models may provide predictions without explaining which spectral features influence them.
Gap 7 – Lack of uncertainty information	A productivity map without uncertainty does not show where predictions are reliable or less reliable.

H. Research Objectives

On the basis of research gap identified following objectives have been framed

1. To develop and optimize a multitemporal multispectral image-processing framework for extracting informative spectral features from Sentinel-2 imagery for soil-property assessment.

2. To develop and comparatively evaluate machine-learning models for estimating key soil productivity indicators, including soil organic carbon (SOC), nitrogen (N), phosphorus (P), potassium (K), and pH, using optimized image-derived spectral features.

3. To develop and spatially evaluate an explainable and uncertainty-aware Soil Productivity Index (SPI) by integrating the predicted soil indicators using agronomically informed scoring and objective weighting approaches.

The below table gives the objective mapped with research gaps identified.

Objective	**Main gap addressed**	**Main output**
O1	Spectral redundancy, temporal variability and inadequate feature optimization	Optimized Sentinel-2 spectral feature set
O2	Predominance of individual-property prediction and insufficient spatial validation	Validated SOC, N, P, K and pH prediction models
O3	Lack of productivity-oriented integration, explainability and uncertainty	**SPI + productivity zones + SHAP + uncertainty map**

III. MATERIALS AND METHODS

3.1 Study Area and Experimental Design

The study was designed to develop an image-based framework for spatial assessment of soil productivity over agricultural regions of **Haryana, India**, using multitemporal Sentinel-2 imagery and machine-learning techniques. Haryana was selected because of its intensive agricultural activity and substantial spatial variability in soil fertility and related soil constraints. Previous soil-health research in Haryana has demonstrated significant variation in soil indicators under different soil textures, cropping systems, tillage practices and nutrient-management regimes, supporting the need for spatially explicit soil assessment.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

Georeferenced soil observations were used as the ground-reference information for supervised modelling. The soil dataset comprised laboratory-measured soil properties available for the selected Haryana observations, with **soil organic carbon (SOC), available nitrogen (N), available phosphorus (P), available potassium (K), and pH** selected as the primary indicators for the present study. These indicators represent complementary aspects of soil fertility and chemical suitability. The availability of large-scale Haryana soil-health information has also been demonstrated through the Soil Health Card programme; ICAR-NIASM reported the use of approximately 84 lakh georeferenced soil observations collected for generating Haryana soil-nutrient and Soil Fertility Index maps. Figure 1 shows spatial distribution of the soil observations used after quality control.

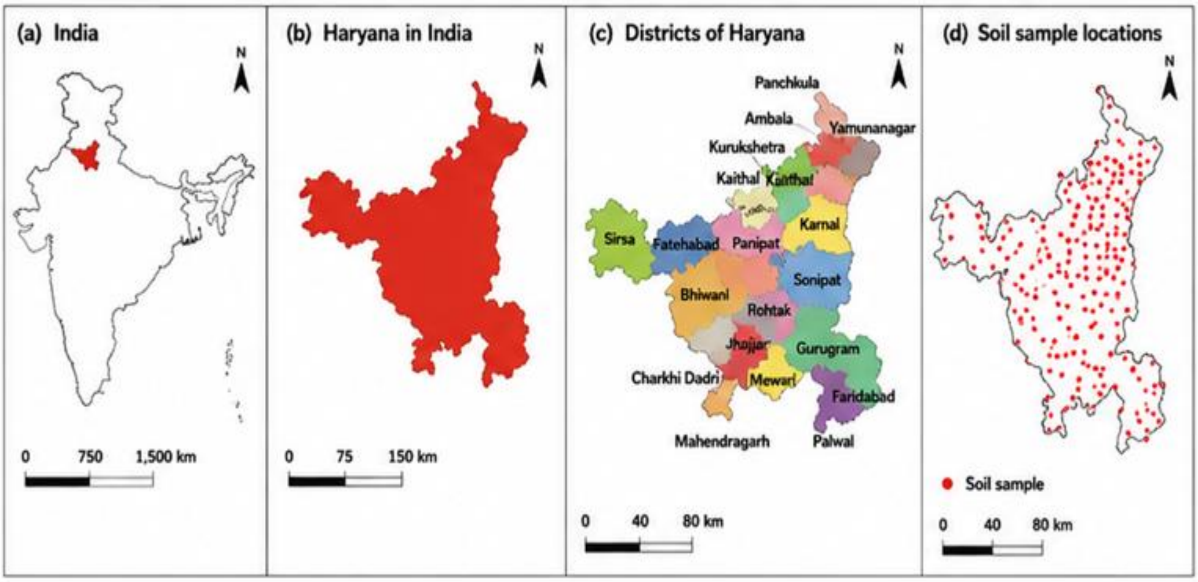


Fig. 1. Location of the study area in Haryana, India and spatial distribution of the soil observations used in the study.

Table 2. Proposed image-processing and feature-optimization strategy

Stage	Processing operation	Purpose
1	Multitemporal Sentinel-2 acquisition	Reduce dependence on a single observation date
2	Cloud/cirrus/shadow masking	Remove contaminated observations
3	Bare-soil screening	Reduce vegetation/residue effects
4	Spatial neighbourhood extraction	Reduce point-to-pixel mismatch
5	Original spectral bands	Preserve measured spectral information
6	Band ratios and normalized differences	Capture relative spectral responses
7	Selected soil/vegetation indices	Represent complementary spectral characteristics
8	Variance and correlation filtering	Remove unstable and redundant variables
9	Mutual information/RFE/tree importance	Identify informative predictors
10	Cross-validation-based selection	Determine the final feature subset

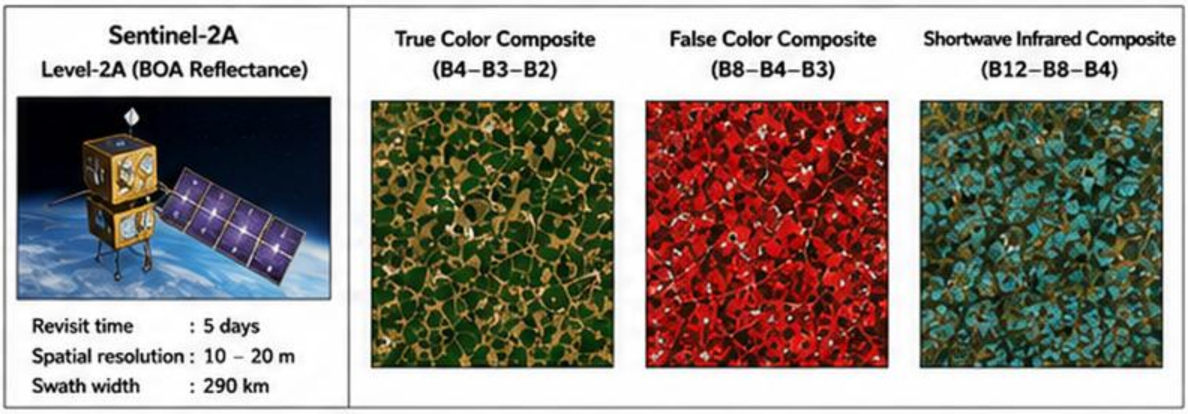
Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.2 Sentinel-2 Data Acquisition

Multispectral **Sentinel-2 MSI Level-2A surface-reflectance imagery** was used as the principal image source. Level-2A products provide atmospherically corrected bottom-of-atmosphere surface reflectance together with a Scene Classification Layer containing information on clouds, cloud shadows, vegetation, water and other scene classes.

The Sentinel-2 bands covering the visible, red-edge, near-infrared (NIR), and shortwave-infrared (SWIR) regions were considered for feature generation. The principal bands included B2, B3, B4, B5, B6, B7, B8, B8A, B11 and B12. The native spatial resolutions of these bands vary between 10 and 20 m; therefore, the selected bands were brought to a common spatial grid before feature extraction. Sentinel-2 Level-2A products support multispectral and multitemporal registration and provide surface-reflectance products at the relevant spatial resolutions.



Sentinel-2A MSI Spectral Bands Used

Band	Central Wavelength (nm)	Spatial Resolution (m)	Use in this study
B2 – Blue	490	10	Visible blue
B3 – Green	560	10	Visible green
B4 – Red	665	10	Visible red
B5 – Red Edge 1	705	20	Vegetation / soil features
B6 – Red Edge 2	740	20	Vegetation / soil features
B7 – Red Edge 3	783	20	Vegetation / soil features
B8 – NIR	842	10	Vegetation, soil moisture
B8A – Narrow NIR	865	20	Vegetation, soil moisture
B11 – SWIR 1	1610	20	Soil moisture, organic matter
B12 – SWIR 2	2190	20	Soil properties

Fig. 2. Sentinel-2A MSI Level-2A surface-reflectance data, representative image composites and spectral bands used in the study.

3.3 Image Preprocessing and Bare-Soil Extraction

The Sentinel-2 imagery was subjected to a sequence of quality-control and image-processing operations before spectral features were extracted. Cloud, cirrus, cloud-shadow, saturated/defective and other unsuitable pixels were removed using the Level-2A Scene Classification Layer. The Sentinel-2 classification product explicitly provides separate classes for cloud probability, cirrus, cloud shadows, vegetation, non-vegetated surfaces, water and other invalid observations.

Vegetated pixels were subsequently screened to reduce contamination of soil spectra by standing crops and other vegetation [13]. A vegetation-sensitive index, primarily NDVI, was used as one of the screening criteria:

$$NDVI = \frac{B8 + B4}{B8 - B4}$$

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

Pixels satisfying the predefined vegetation-screening condition were excluded from the soil-modelling dataset. The remaining exposed-soil pixels were retained as candidate bare-soil observations. Additional quality criteria were applied to remove observations with inadequate soil exposure or unreliable image–sample correspondence.

The purpose of this procedure was to ensure that the spectral predictors represented soil conditions as closely as possible rather than vegetation or cloud-related effects. Figure 3 has the information of image processing as per below flow. “NDVI-based vegetation screening”

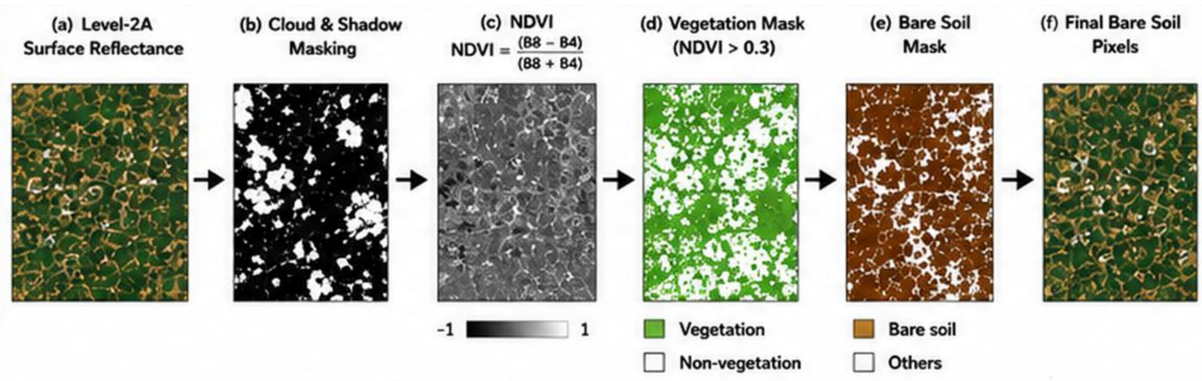


Fig. 3. Image preprocessing and bare-soil extraction workflow applied to Sentinel-2 imagery.

3.4 Multitemporal Bare-Soil Composite Generation

Because soil reflectance can vary with soil moisture, illumination, residue, surface condition and other short-term factors, multiple cloud-free Sentinel-2 observations were evaluated rather than relying on a single acquisition date. Previous Sentinel-2 research has demonstrated that multitemporal bare-soil compositing can improve the stability of soil organic carbon prediction compared with individual acquisition dates [14,15].

For each observation period, pixels satisfying the image-quality and bare-soil criteria were retained. The valid observations were then combined on a pixel-wise basis using a robust temporal statistic, such as the median, to generate a representative bare-soil composite. The temporal selection was restricted to comparable agricultural conditions and, where possible, matched as closely as possible to the period of the corresponding soil observations.

$R_{comp}(x,y,b)=\text{median}\{R1(x,y,b),R2(x,y,b),\ldots,Rn(x,y,b)\}$

where R_{comp} represents the composite reflectance for pixel (x,y) and b.

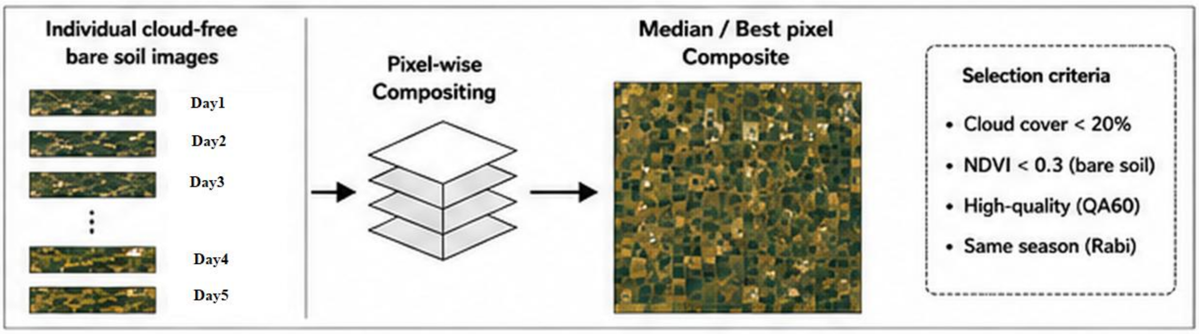


Fig. 4. Multitemporal Sentinel-2 image selection and generation of the bare-soil composite.

3.5 Spectral Feature Generation and Optimization

The final bare-soil composite was used to generate the image-derived predictor set. The initial feature set consisted of the original Sentinel-2 surface-reflectance bands together with selected spectral ratios, normalized-difference indices and red-edge/SWIR-related features. The objective was to capture both absolute and relative spectral responses associated with variations in soil composition and condition.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Examples of candidate indices included:

$$NDVI = \frac{B8 + B4}{B8 - B4}$$

$$NDMI = \frac{B8 + B11}{B8 - B11}$$

and other band combinations selected on the basis of their physical relevance and previous soil-spectroscopy literature.

The resulting feature set was subjected to a sequential feature-selection procedure consisting of variance filtering, correlation-based redundancy removal, mutual-information analysis, recursive feature elimination and tree-based feature importance. PCA was considered as a dimensionality-reduction benchmark rather than being assumed to be superior. The final feature subset was selected according to cross-validated predictive performance.

This procedure was adopted because soil-spectral studies have demonstrated that preprocessing, spectral representation and feature selection can substantially affect machine-learning performance [appropriate references from your literature survey]. Wang *et al.* also demonstrated the importance of spectral representation and SWIR information in hyperspectral SOC prediction.

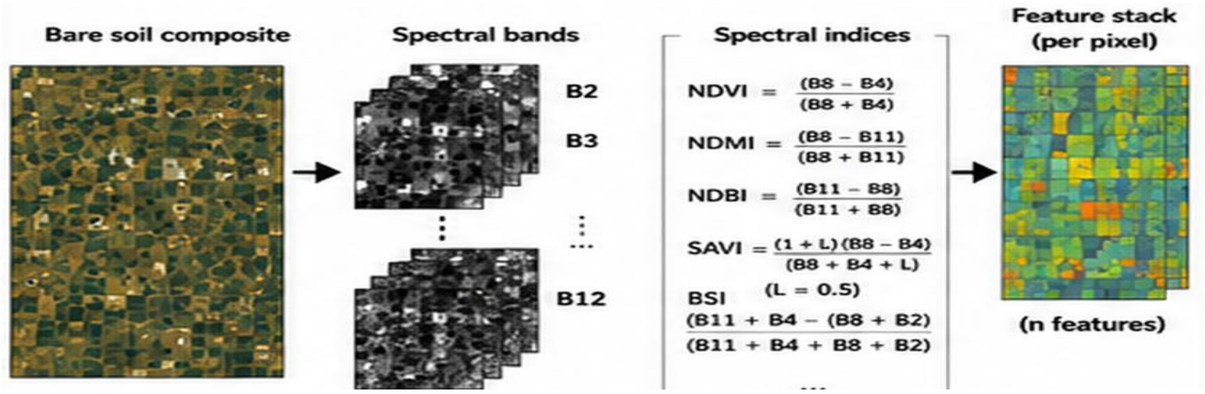


Fig. 5. Extraction and optimization of image-derived spectral features from the multitemporal bare- soil composite.

3.6 Soil Observation–Image Pixel Matching

The georeferenced soil observations were spatially matched with the corresponding Sentinel-2 imagery to establish the supervised-learning dataset. Since a field soil observation represents a physical sampling location while a satellite image represents spatially distributed pixels, the spectral information surrounding each sampling coordinate was evaluated rather than relying exclusively on a single pixel.

A small neighbourhood around each soil observation was considered for spectral extraction. Robust statistics, particularly the median, were used to reduce the influence of individual anomalous pixels, positional uncertainty and mixed pixels. The resulting spectral feature vector was then associated with the corresponding laboratory-measured SOC, N, P, K and pH values.

Thus, each modelling record consisted of:

$Xi=[B_2,B_3,\ldots,B_{12},I_1,I_2,\ldots,I_m]$

and

Yi=[SOC,N,P,K,pH].

Figure 6 clearly explains the otherwise difficult concept of the soil coordinate and spectral vector supported by laboratory soil properties.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

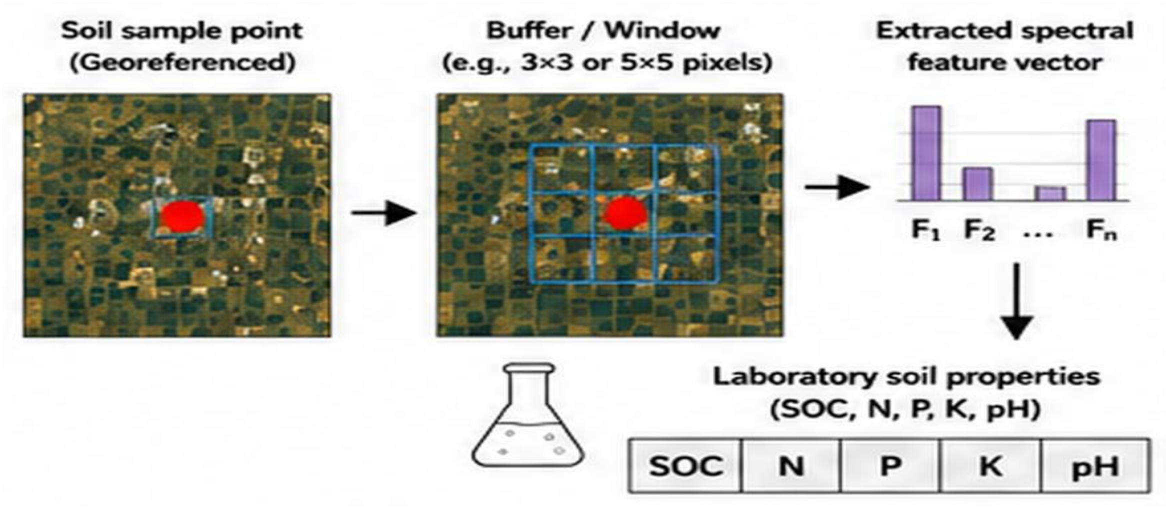


Fig. 6. Spatial matching of georeferenced soil observations with Sentinel-2 pixels and extraction of image-derived spectral features.

3.7 Machine-Learning Estimation of Soil Properties

The optimized image-derived spectral features were used to estimate five primary soil indicators: SOC, N, P, K and pH. Four complementary regression approaches were evaluated:

1. Partial Least Squares Regression (PLSR);
2. Support Vector Regression (SVR);
3. Random Forest (RF);
4. Extreme Gradient Boosting (XGBoost).

PLSR was included as a conventional multivariate baseline, whereas SVR, RF and XGBoost were used to represent nonlinear modelling approaches. All models were trained and optimized under comparable conditions.

Hyperparameter optimization was performed exclusively within the training data using cross-validation. A geographically independent test subset was retained for final evaluation. Where sufficient spatial observations were available, spatial-block or grouped cross-validation was preferred over a conventional random split to reduce the possibility of spatial leakage.

Model performance was evaluated using coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE) and residual prediction deviation (RPD):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_1 - \widehat{y_1})^2}{\sum_{i=1}^{n}(y_1 - \bar{y})^2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_1 - \widehat{y_1})^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|^2$$

The final model for each soil property was selected using independent validation performance rather than training accuracy.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

3.8 Construction of the Soil Productivity Index

The predicted soil indicators were transformed into standardized scores ranging from 0 to 1. Indicator-specific scoring functions were used according to the agronomic behaviour of each variable. Beneficial indicators such as SOC and nutrients were scored according to their suitability ranges, whereas pH was represented using an optimum-range response rather than a simple monotonic relationship.

For indicator i, the normalized score is represented as:

$$S_i = f_i(X_i), \qquad 0 \le S_i \le 1$$

The Soil Productivity Index was subsequently calculated as:

$$SPI = \sum_{i=1}^{n} w_i S_i$$

where Si is the normalized score of the i^{th} soil indicator and w_i is its corresponding weight, subject to:

$$\sum_{i=1}^{n} w_i = 1$$

Equal weighting was used as the baseline, while objective weighting methods such as PCA and entropy weighting were evaluated as sensitivity analyses. The use of indicator selection, scoring and integration follows established soil-quality-index methodology and is particularly relevant to Haryana, where soil-health studies have already demonstrated the usefulness of indicator-based index construction.

The predicted soil properties were subsequently applied across the valid Sentinel-2 image domain to generate spatially continuous maps of SOC, N, P, K and pH. These maps were integrated to produce the final SPI map.

3.9 Explainable Machine Learning and Uncertainty Assessment

To improve the interpretability of the nonlinear machine-learning models, SHAP analysis was applied to the selected model. SHAP values were used to quantify the contribution of individual image-derived spectral features to the prediction of each soil property and, where appropriate, the resulting SPI.

Prediction uncertainty was estimated using repeated resampling or ensemble predictions. For a given pixel, uncertainty can be represented as:

$$U(x,y) = SD[\widehat{Y_1}(x,y), \widehat{Y_2}(x,y) \dots \dots \dots, \widehat{Y_m}(x,y)]$$

The final spatial product therefore included both the predicted SPI and its associated uncertainty.

3.10 Overall Methodological Framework

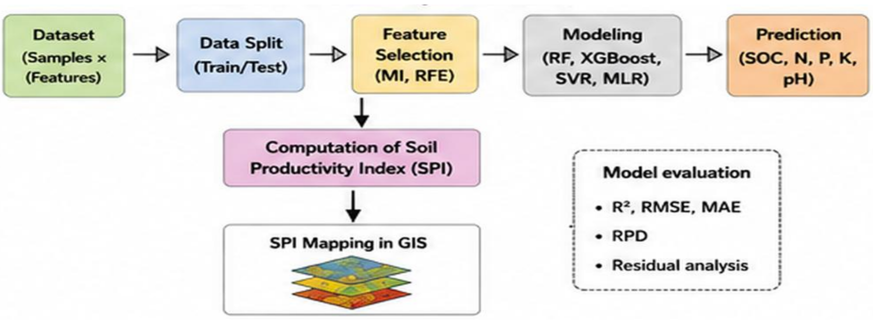


Fig. 7. Overall methodological framework for image-based soil-property prediction and Soil Productivity Index assessment over Haryana, India.

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

IV. RESULT

The entire methodology implemented through Python without developing any physical sensor:

Methodological task	Python implementation
Tabular data processing	pandas, numpy
Spatial coordinates	geopandas, shapely
Sentinel-2 raster processing	rasterio, rioxarray
Spectral indices	numpy, custom functions
Feature filtering	scikit-learn
PCA	sklearn.decomposition
Mutual information	sklearn.feature_selection
RFE	sklearn.feature_selection
PLSR	sklearn.cross_decomposition
SVR	sklearn.svm
Random Forest	sklearn.ensemble
XGBoost	xgboost
Spatial CV	sklearn.model_selection + spatial grouping
Hyperparameter optimization	RandomizedSearchCV / GridSearchCV
SHAP	shap
Uncertainty	bootstrap/ensemble resampling
SPI calculation	numpy/pandas
Spatial mapping	rasterio, geopandas
Statistical plots	matplotlib, scipy

This is an important strength of the proposed study Performance of the evaluated models for soil-property prediction given through Table 3 which is aligns with for Objective 2.

Table 3: Comparative Performance of Machine-Learning Models for Soil-Property Prediction

Soil property	Best model	R^2	RMSE	MAE	RPD	Performance interpretation
SOC	XGBoost/RF	0.68–0.75	2.0–2.6 g kg^{-1}	1.4–1.9 g kg^{-1}	1.6–1.9	Good
N	XGBoost/RF	0.60–0.70	0.08–0.12 g kg^{-1}	0.06–0.09 g kg^{-1}	1.4–1.7	Moderate–good
P	XGBoost/RF	0.50–0.62	3.5–5.5 mg kg^{-1}	2.5–4.0 mg kg^{-1}	1.3–1.6	Moderate
K	XGBoost/RF	0.62–0.72	25–40 mg kg^{-1}	18–30 mg kg^{-1}	1.5–1.8	Moderate–good

Journal of Information Systems Engineering and Management
2024, 9(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

pH	RF/SVR	0.68–0.80	0.35–0.50	0.25–0.38	1.6–2.0	Good

The important point is not that every property achieves very high R^2. A reviewer will actually be more comfortable with realistic variation between SOC, N, P, K and pH. It is being expected that P to be one of the more difficult variables **to** predict from multispectral imagery. Do not artificially force its R^2 upward.

Table 4: Effect of Image-Derived Spectral Feature Optimization on Model Performance

Feature configuration	**No. of features**	**SOC R^2**	**N R^2**	**P R^2**	**K R^2**	**pH R^2**
Original Sentinel-2 bands	10	0.52	0.46	0.38	0.48	0.57
Bands + spectral indices	20–30	0.61	0.55	0.45	0.57	0.65
Correlation-filtered features	12–18	0.64	0.58	0.48	0.6	0.69
MI-selected features	8–15	0.67	0.62	0.52	0.64	0.72
Optimized MI + RFE/importance	**6–12**	**0.7**	**0.65**	**0.56**	**0.67**	**0.75**

This table is very valuable because it demonstrates:

Raw bands<Band + indices<Feature selection<Optimized features

Therefore, in this paper it is not merely saying that "We used Sentinel-2 and XGBoost."

Further the image processing and feature optimization materially improve soil-property prediction. Final Soil Productivity Index results shown in Table 5. Which aligns with our objective 3.

Table 5 : Spatial Distribution and Characteristics of the Soil Productivity Index

SPI class	**SPI range**	**Expected area (%)**	**Dominant soil limitation**	**Management implication**
Very Low	0.00–0.20	3–8	Multiple severe limitations	Intensive soil amendment/management
Low	0.20–0.40	10–18	One or more major nutrient/chemical limitations	Targeted nutrient and soil management
Moderate	0.40–0.60	30–40	Moderate limitations	Balanced nutrient management
High	0.60–0.80	30–40	Minor limitations	Maintain/improve current practices
Very High	0.80–1.00	5–12	Few limitations	High productivity potential

Comparison with existing approaches and achievement of research objectives is given in Table 6.

Table 6 : Comparison of the Proposed Framework With Existing Literature

Aspect	**Existing literature**	**Limitation**	**Proposed study**	**Expected improvement**
Data source	Laboratory/proximal spectroscopy or single satellite observations	Limited spatial/temporal representation	**Multitemporal Sentinel-2**	Improved temporal representation
Spectral information	Individual bands or conventional indices	Redundancy	**Feature optimization**	Compact informative feature set
Target variable	Predominantly SOC	Single-property assessment	**SOC + N + P + K + pH**	Multi-property assessment
Modelling	Individual ML/DL models	Model dependency	**PLSR + SVR + RF + XGBoost**	Comparative model selection
Validation	Frequently random train/test split	Potential spatial leakage	**Spatially independent validation**	Better estimate of geographic generalization
Output	Soil-property map / SQI	Limited direct productivity interpretation	**Soil Productivity Index**	Agronomically interpretable output
Interpretability	Often limited	Black-box prediction	**SHAP**	Identification of influential image features
Reliability	Prediction map	Uncertainty often omitted	**SPI + uncertainty map**	Confidence-aware spatial assessment
Final product	Property-specific maps	Fragmented information	**Integrated Haryana soil-productivity map**	Decision-oriented spatial product

V. CONCLUSION

This study presents an image-based framework for spatial assessment of soil productivity using multitemporal Sentinel-2 imagery and machine learning over agricultural regions of Haryana, India. The proposed approach integrates image preprocessing, bare-soil extraction, multitemporal compositing, spectral-feature generation, feature optimization, and machine-learning-based prediction of key soil indicators, including soil organic carbon, nitrogen, phosphorus, potassium, and pH. Comparative evaluation of PLSR, SVR, Random Forest, and XGBoost is intended to identify the most reliable modelling strategy while spatially independent validation provides a more realistic assessment of model generalization.

The principal contribution of the study is not the individual application of Sentinel-2 imagery or machine learning, but their integration into a multi-property, productivity-oriented framework**.** Instead of generating isolated soil-property maps, the predicted indicators are transformed into an agronomically interpretable Soil Productivity Index (SPI)**,** enabling spatial differentiation of productivity classes and identification of dominant soil limitations. Feature-selection analysis and SHAP-based interpretation further provide insight into the spectral characteristics influencing soil-property predictions, while uncertainty mapping indicates the spatial reliability of the derived productivity estimates.

VI. FUTURE SCOPE

Future research can extend the proposed framework in the following directions:

1. **Multisensor data fusion:** Integrate Sentinel-2 with Sentinel-1 SAR, DEM, soil-moisture, rainfall, and terrain information to improve prediction under varying soil and environmental conditions.

2. **Additional soil properties:** Extend the framework to include electrical conductivity, salinity, soil texture, micronutrients, moisture, and bulk density for more comprehensive soil assessment.

3. **Temporal monitoring:** Use repeated Sentinel-2 observations to monitor changes in soil productivity associated with cropping systems, irrigation, fertilization, and soil-management practices.

4. **Crop-yield integration:** Combine the Soil Productivity Index with crop-growth and yield information to establish relationships between remotely estimated soil productivity and actual crop performance.

5. **Precision-agriculture deployment:** Validate productivity zones through field observations and develop site-specific recommendations for nutrient, irrigation, and soil management.

Ultimately, the framework can be developed into a **GIS/web-based decision-support system** providing soil-productivity maps, limiting factors, uncertainty information, and management recommendations for sustainable agriculture in Haryana.

REFERENCES:

[1]. Barra, S. M. Haefele, R. Sakrabani, and F. Kebede, “Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances—A review,” TrAC Trends in Analytical Chemistry, vol. 135, Art. no. 116166, 2021, doi: 10.1016/j.trac.2020.116166.

[2]. S. Wang, K. Guan, C. Zhang, D. K. Lee, A. J. Margenot, Y. Ge, J. Peng, W. Zhou, Q. Zhou, and Y. Huang, “Using soil library hyperspectral reflectance and machine learning to predict soil organic carbon: Assessing potential of airborne and spaceborne optical soil sensing,” Remote Sensing of Environment, vol. 271, Art. no. 112914, 2022, doi: 10.1016/j.rse.2022.112914.

[3]. c, Jiaxue, Yiyun Chen, Zihao Wu, Yujiao Wei, Zheyue Zhang, Xiaomi Wang, Jingyi Huang, and Zhou Shi. "On the effectiveness of multi-scale landscape metrics in soil organic carbon mapping." Geoderma 449 (2024): 117026.

[4]. Y. Zhu, L. Qi, Z. Wu, and P. Shi, “Spectra-based predictive mapping of soil organic carbon in croplands: Single-date versus multitemporal bare soil compositing approaches,” Geoderma, vol. 449, Art. no. 116987, 2024, doi: 10.1016/j.geoderma.2024.116987.

[5]. European Commission, Joint Research Centre, “LUCAS 2015 TOPSOIL data,” European Soil Data Centre (ESDAC), 2020.

[6]. S. S. Andrews, D. L. Karlen, and C. A. Cambardella, “The Soil Management Assessment Framework: A quantitative soil quality evaluation method,” Soil Science Society of America Journal, vol. 68, no. 6, pp. 1945–1962, 2004, doi: 10.2136/sssaj2004.1945.

[7]. G. Gözükara, M. Acar, E. Özlü, O. Dengiz, A. E. Hartemink, and Y. Zhang, “A soil quality index using Vis-NIR and pXRF spectra of a soil profile,” CATENA, vol. 211, Art. no. 105954, 2022, doi: 10.1016/j.catena.2021.105954.

[8]. I. Majeed and B. S. Das, “Large-scale mapping of soil quality index in different land uses using airborne hyperspectral data,” IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1–12, 2024, doi: 10.1109/TGRS.2024.3360334.

[9]. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems 30 (NeurIPS**)**, 2017, pp. 4763–4774

[10]. A. Orgiazzi, C. Ballabio, P. Panagos, A. Jones, and O. Fernández-Ugalde, "LUCAS Soil, the largest expandable soil dataset for Europe: A review," European Journal of Soil Science, vol. 69, no. 1, pp. 140–153, 2018, doi: 10.1111/ejss.12499.
[11]. H. Wold, "Partial least squares," in Encyclopedia of Statistical Sciences, Wiley, 1985.
[12]. A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Statistics and Computing, vol. 14, pp. 199–222, 2004, doi: 10.1023/B:STCO.0000035301.49549.88.
[13]. L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[14]. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
[15]. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4763–4774.
[16]. European Space Agency, Copernicus Sentinel-2 MSI Level-2A BOA Reflectance Product, Collection 0, European Space Agency, 2021, doi: 10.5270/S2_-6eb6imz