

Machine Learning-Based Classification of Soil Reflectance Signatures Acquired Using ASD FieldSpec Spectroscopy

Neha Baban Chavan¹ and Dr. Sonali B. Kulkarni²

Ph. D Scholar, Department of Computer Science and IT¹

Associate Professor, Department of Computer Science and IT²

Dr. Babasaheb Ambedkar Marathwada University Chhatrapati Sambhajnagar, India

Abstract: *The proper classification of soils is crucial for sustainable agriculture and efficient land resource management. Traditional methods of soil analysis in laboratory conditions are reliable but are laborious, time-consuming, and destructive. Spectral reflectance measurement with hyperspectral spectroscopy is an alternative approach that allows one to receive fast and non-destructive results. In this research, the spectral signature was recorded from 16 soil samples using the ASD FieldSpec spectroradiometer within the 350-2500 nm spectral range. Each soil sample was scanned 10 times; thus, the data set included 160 soil hyperspectral spectral signatures that contain 2151 features. The obtained data was pre-processed, and Principal Component Analysis (PCA) was used to reduce the number of features from 2151 to 50 while keeping the main spectral features of the soil. Two supervised machine learning classification algorithms, i.e., Random Forest and Support Vector Machine (SVM), were used. The classifiers demonstrated 100% accuracy; precision, recall, and F1-score were equal to 1.00. Confusion matrix did not show any misclassification. Fivefold cross-validation demonstrated the mean accuracy of 100%. The results confirm that hyperspectral reflectance spectroscopy together with machine learning constitutes a robust, efficient, and non-destructive method for classifying the spectral signatures of soil and has great application in precision agriculture and environmental monitoring.*

Keywords: Soil Classification, Machine Learning, ASD FieldSpec Spectroscopy, Spectral Signature Analysis, Feature Extraction..

I. INTRODUCTION

Soils are an important part of agricultural production, environmental assessments, ecology preservation, and land management. Effective soil analysis is very important for making correct decisions, increasing production efficiency, and sustainable land management [3, 6]. Conventional methods of soil analysis are often expensive, time-consuming, laborious, destructive, and need a lot of preparation. Due to these restrictions, many rapid and non-destructive methods of soil analysis were developed [1, 14].

Spectroscopy gives detailed information in different ranges of spectrum and, therefore, can be used for characterizing soils. Spectral sensors allow acquiring hundreds to thousands of continuous spectral bands that help determine small differences in soil reflectance associated with their physical and chemical characteristics [3, 8]. VIS, NIR, and SWIR regions give useful information about the soil composition, minerals, humidity, and organic matter. As a result, hyperspectral reflectance spectroscopy has been widely used in agriculture, geology, and environmental studies [17, 20].

The ASD FieldSpec spectrometer is famous due to its high spectral resolution and wide wavelength range. The device provides accurate measurement of reflectance under laboratory conditions and is commonly used for creation of



spectral libraries, soil reflectance, minerals, vegetation research and calibration of remote sensing measurements [3, 6]. High-resolution information obtained by the ASD FieldSpec spectrometer helps to discriminate between the spectral characteristics of soil and, therefore, allows effective soil characterization and classification [10, 15].

Nevertheless, hyperspectral spectroscopy produces high-dimensional data sets. There may be thousands of bands in one spectrum, and they are highly correlated or redundant, complicating computations and decreasing the effectiveness of machine learning algorithms [2, 4]. Therefore, dimensionality reduction methods like PCA are often used to decrease the number of correlated or redundant bands [5, 7].

Advances in artificial intelligence and machine learning made a revolution in hyperspectral data analysis. Machine learning algorithms can find spectral patterns and classify reflectance spectra [2], [4]. These algorithms include the RF and SVM which have been studied extensively due to their potential to handle high-dimensional spectral data and provide a reliable classification performance [5], [11].

In this study, hyperspectral reflectance signatures are obtained from 16 soil sampling locations inside the Dr. Babasaheb Ambedkar Marathwada University (BAMU) campus. Ten repeated spectra are obtained for each soil sample. The dataset was reduced by applying principal component analysis. The spectral data is analyzed using the Random Forest classifier and classifies the 16 soil sampling locations.

The primary goal of this research is to analyze the soil spectral signatures using the hyperspectral reflectance spectroscopy along with machine learning. The proposed approach is aimed at developing an accurate and non-destructive method to distinguish soil spectral signatures taken from different sampling locations.

Contributions of the Study

The contributions of the study can be summarized as follows:

- * Development of a machine learning approach for analysis of hyperspectral soil reflectance signature using ASD FieldSpec spectroscopy and Random Forest classifier.
- * Collection of a dataset containing 160 hyperspectral soil signatures from 16 sampling locations within the BAMU campus where each soil sample is measured through 10 repeated spectral scans.
- * Dimensionality reduction of dimensional hyperspectral data using Principal Component Analysis (PCA) reducing the original 2051 spectral features to 50 principal components while preserving the dominant spectral information.
- * The Random Forest model was evaluated for performance and robustness in terms of accuracy, precision, recall, F1 – score, confusion matrix and 5-fold cross-validation.
- * The proposed method provides reliable and non-destructive analysis of soil spectral signatures by combining hyperspectral reflectance spectroscopy with machine learning.

The remainder of this paper is organized as follows. Section II reviews literature related to hyperspectral soil analysis and machine learning-based soil classification. Section III presents the materials and methods including data acquisition, preprocessing, feature extraction, machine learning model development and evaluation metrics. Section IV discusses the results and performance evaluation. Finally Section V concludes the study. Outlines future research directions.

II. RELATED WORK

The analysis of soil is quite an essential procedure in agriculture, environment monitoring, and land resource management. In recent years, a variety of sensing techniques, hyperspectral spectroscopy, and machine learning approaches were studied for fast and non-invasive soil characterization.

It is possible to apply machine learning classification of soil reflectance spectra, obtained via ASD FieldSpec spectroscopy, but the efficiency of this approach greatly depends on pre-processing, feature selection, sensor parameters, and the target for classification, not prediction [8][19][15][18]. According to scientific literature, ASD FieldSpec technology is applied for either soil horizon, suborder, contamination class, and type classification, or for the discrimination of soil parameters via regression technique ([8][9][13][20].



Classification Performance For soil horizons and suborders, the ASD FieldSpec-based machine learning models performed very well, with neural networks exhibiting 97% accuracy in classification of horizon and suborder classification accuracies being 95% to 98% [8]. Another study involving horizon but using hyperspectral imaging rather than non-imaging ASD data still obtained high accuracies but not as good as the former case, where the highest was 89.9% in VIS-NIR and 86.5% in SWIR with the help of random forest classifier, proving the possibility of horizon boundaries classification despite spectral overlap [9]. For soil type classification, ASD FieldSpec 4 spectra collected in the field along with machine learning classified three soil types with more than 95% accuracy in a precision agriculture study [19]. A similar study involving hyperspectral data used spectral libraries and machine learning with satellite imagery to classify three soil types with over 95% classification accuracy [20]. For contamination classification, vis-NIR spectroscopy and machine learning used support vector classifiers to distinguish between high and low arsenic contamination across Europe with AUC of 0.89 and accuracy 0.83 after tuning, and the addition of an ensemble to the analysis slightly improved AUC to 0.90 [13]. In an earlier study on soil arsenic diagnosis, the optimal overall accuracies achieved for the different methods were 86% to 89% across RF, ANN, and SVMs, without any significant difference in their performance [18].

Soil reflectance classification is possible since the spectra contain information regarding the composition of the soil, such as clays, iron, carbonates, soil organic matter, and moisture, as well as unique absorption bands at the 400-2500 nm range [7]. Discrimination performance has been attributed to the particle size, soil organic carbon, Fe oxides, and carbonates in various studies, including regions of arsenic classification at 471 and 2422 nm and carbonate sensitive region at 2350 nm [7][8][9][13][14].

Dimensionality is an important challenge in this case, and several works show that pre-processing and feature selection help increase either efficiency or accuracy [7][18].

TABLE I: Representative performance trends of machine-learning algorithms found for the problem of soil spectral classification or soil spectroscopy using ASD technology.

Task	Best-performing model(s) reported	Key Result
Horizons / suborders	NN, RF	97% horizon accuracy; 95-98% suborders [8]
Soil types	ML with ASD FieldSpec 4	>95% success for three soil types [19]
Arsenic Classes	SVC, ensemble	AUC 0.89-0.90, accuracy 0.83 [13]
Texture / attributes	KNN, RF, NN, CNN, MBL	Nonlinear models often beat PLSR [10][11][17]

In the case of similar ASD or soil spectroscopy problems, deep learning and nonlinear approaches often outperform linear baselines, such as CNN vs PLSR, MBL vs PLSR/SVMR/BRT, SVMR vs PLSR for nitrogen-related properties, and RF wrappers vs standalone RF or PLSR with feature selection [7][11][12][16]. However, there is not enough information on the existence of one classifier being better than others since the top classifier depends on sample type, range of spectra, pre-processing steps and the goal of the study (horizons, contamination, texture, or soil properties) [7][14][17].

III. RESEARCH GAP

From the literature review, it is clear that the past researches mainly concentrated on assessing soil quality, estimating soil moisture, mapping soil salinity and predicting soil properties using hyperspectral spectroscopy along with machine learning algorithms [3], [6], [17]. Many researchers have also used spectral data to detect the presence of heavy metals in soil, analyze soil contamination and precision agriculture applications [12], [13], [18]. While many machine learning algorithms and dimensionality reduction methods have been commonly applied to hyperspectral data analysis, there has not been much work done regarding the classification of soil reflectance spectra obtained by ASD FieldSpec spectroscopy in controlled laboratory environment [2], [4], [11]. In addition to that, only a few studies have studied the combination of Principal Component Analysis (PCA) with Random Forest (RF) and Support Vector Machine (SVM) classifiers for repeated soil hyperspectral spectra from multiple sampling sites [5], [7], [20]. Hence, in order to fill these research gaps, this current research attempts to develop a machine learning framework that makes use of ASD



FieldSpec hyperspectral reflectance spectroscopy, Principal Component Analysis (PCA), and Random Forest (RF) and Support Vector Machine (SVM) classifiers.

IV. MATERIALS AND METHODS

The proposed methodology was created for classification of soil spectral signatures obtained by means of hyperspectral reflectance spectroscopy. The complete workflow consists of seven major stages: (i) study area selection, (ii) soil sample collection, (iii) experimental work and hyperspectral data acquisition, (iv) spectral data preprocessing, (v) dimensionality reduction using Principal Component Analysis (PCA), (vi) soil spectral signature classification using Random Forest (RF) and Support Vector Machine (SVM), and (vii) performance evaluation. The overall workflow is illustrated in Fig. 1.

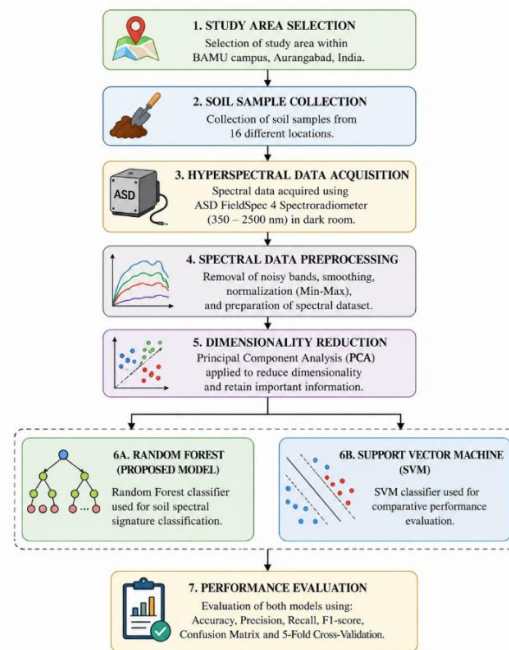


Fig. 1 Workflow of the proposed methodology for hyperspectral soil spectral signature classification.

Study Area

The study was conducted within the campus of Dr. Babasaheb Ambedkar Marathwada University (BAMU), Chhatrapati Sambhajinagar, Maharashtra, India. The university campus comprises diverse land cover conditions, including academic buildings, botanical gardens, residential areas, open grounds, roadside locations, and vegetated regions. Sixteen representative sampling locations were selected across the campus to capture the spatial variability of soil reflectance characteristics. Each sampling location was assigned a unique identification label (S1–S16) for further spectral analysis.



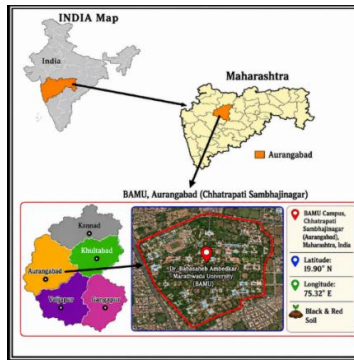


Fig. 2 The Geographical Location Map of the Study Area

Soil Sample Collection

Soil samples from the sixteen selected locations were collected under dry weather conditions. These samples were taken to the laboratory where the stones, plants and other impurities were removed. Further, the samples were air dried in ambient laboratory conditions and were homogenized to provide a uniform surface for spectral measurements. Thus, the impact of external factors on the results was decreased.

ASD FieldSpec Spectral Data Acquisition

Hyperspectral reflectance spectra of samples were measured with use of ASD FieldSpec spectrometer for wavelengths from 350 nm to 2500 nm that includes VIS, NIR, SWIR regions.

In order to decrease random instrumental errors and increase reliability of measurement, ten repeated spectral scans were performed for each sample in the laboratory conditions. Thus, 160 hyperspectral spectral signatures were obtained.

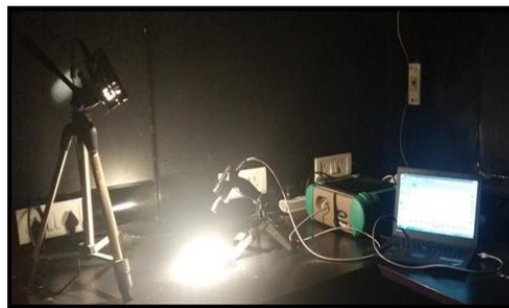


Fig. 3 Field spec4 Spectroradiometer setup for Soil Spectral measurement

The ASD FieldSpec spectrometer captures reflectance data points continuously within the spectral range, generating multidimensional spectral signatures that show the interaction between the electromagnetic radiation and the soil surface.

Experimental Work

The experimental part was performed in laboratory settings using an ASD FieldSpec spectroradiometer, which performs measurements of the spectral reflectance within the 350-2500 nm spectral range. The selected spectral range includes the Visible (VIS), Near-Infrared (NIR), and Short-Wave Infrared (SWIR) wavelengths that include valuable information about soil minerals, organic content, and moisture.

Prior to the measurement process, the instrument was calibrated using the certified white reference panel. The use of white reference calibration makes it possible to eliminate the effect of external factors on the measurements results.



The soil sample was placed into the clean sample holder with uniform illumination. The optical probe of the ASD FieldSpec was located vertically above the soil surface maintaining the same distance during the entire experiment. To enhance the accuracy of measurements, each soil sample was measured ten times and corresponding spectra were saved for the subsequent processing.

In total, 16 different locations for soil sampling were considered in the study. Because each location was analyzed ten times, there were 160 spectral signatures in the final dataset.

The spectral reflectance was determined based on the ratio of the reflected radiance from the soil sample and reflected radiance from the white reference panel:

$$R(\lambda) = \frac{L_s(\lambda)}{L_r(\lambda)} \quad \text{-----} (1)$$

where

- $R(\lambda)$ - reflectance at wavelength λ ,
- $L_s(\lambda)$ - reflected radiance from the soil sample,
- $L_r(\lambda)$ - reflected radiance from the white reference.

Thus, the obtained reflectance spectra were used as the input data for machine learning.

Spectral Data Preprocessing

The hyperspectral dataset acquired from the ASD FieldSpec included 2151 wavelength bands for each spectral signature. The following preparatory steps were undertaken before classification of the data.

First, wavelength bands were extracted and used as feature names. Spectral signatures were placed in rows to form the feature matrix.

The hyperspectral dataset is presented by the formula:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad \text{-----} (2)$$

where

- $m=160$ - the number of spectral signatures,
- $n=2151$ - the number of spectral bands.

Class labels for each sampling location in the form of

$$y = [y_1, y_2, \dots, y_m]^T \quad \text{-----} (3)$$

were identified for each sample where each class corresponds to one soil sampling location.

Once the dataset was preprocessed, it was ready to be subjected to PCA dimensionality reduction.

Principal Component Analysis (PCA)

Hyperspectral datasets usually comprise of several hundreds or even thousands of spectral bands most of which are highly correlated. High-dimensional datasets result in increased computational complexity and reduced classification accuracy. Therefore, a dimensionality reduction technique, namely Principal Component Analysis (PCA), was applied to retain maximum amount of spectral information by reducing dimensionality of the dataset.

First, the mean-centered dataset was calculated:

$$X_c = X - 1\mu^T \quad \text{-----} (4)$$

Where

$$\mu_j = \frac{1}{m} \sum_{i=1}^m X_{ij} \quad \text{-----} (5)$$

Next, the covariance matrix was calculated in the following way:

$$C = \frac{1}{m-1} X_c^T X_c \quad \text{-----} (6)$$



The principal components were calculated by solving the eigenvalue problem:

$$Cv = \lambda v \quad \text{-----}(7)$$

where

- v - eigenvector,
- λ - eigenvalue.

The feature space transformation was calculated as follows:

$$Z = XW \quad \text{-----}(8)$$

where

- W - matrix comprising of selected eigenvectors,
- Z - transformed feature matrix.

In this study, PCA reduced the dimensionality of the feature space from 2151 spectral bands to 50 principal components.

Machine Learning-Based Soil Spectral Signature Classification

The dataset after PCA dimensionality reduction was classified using two supervised machine learning algorithms:

- Random Forest (RF)
- Support Vector Machine (SVM)

Dataset was split into 80% of training data and 20% testing data with the help of stratified sampling.

G.1 Random Forest Classifier

Random Forest is an ensemble method where many decision trees are built and combined in order to get a final prediction based on majority voting.

For a certain input x , the final output class of the model will be

$$\hat{y}(x) = \arg \max_{c \in \mathcal{C}} \sum_{i=1}^k I(T_i(x)=c) \quad \text{-----}(9)$$

where

- $T_i(x)$ - the prediction of the i th decision tree,
- $k=300$ - the total number of decision trees.

During the training stage of the tree, a bootstrap sample from the training set is used and random features are selected for each split.

Random Forest algorithm was used with the following parameters:

- Number of trees = 300
- Random state = 42
- Stratified train-test split = 80:20

and this model was referred to as the proposed classifier.

G.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised classifier algorithm which finds the optimal hyperplane that separates the classes by maximizing the margin.

Decision function is defined as

$$f(x) = w^T x + b \quad \text{-----}(10)$$

where

- w - weight vector,
- b - bias.



Optimization problem can be formulated as

$$\min_{w,b,\xi} \frac{1}{2} |w|^2 + C_s \sum_{i=1}^m \xi_i \quad \text{-----(11)}$$

subject to

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \text{-----(12)}$$

where

- y_i - class label,
- x_i - training sample.

SVM was used as a benchmark classifier in order to test the efficiency of the proposed Random Forest model.

Performance Evaluation

The performances of both Random Forest and SVM were evaluated based on common performance criteria.

Accuracy

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{-----(13)}$$

Precision

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{-----(14)}$$

Recall

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{-----(15)}$$

F1-score

$$F1 = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{-----(16)}$$

Where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively.

In addition, in order to measure the robustness of the models, a five-fold cross validation procedure was conducted. The data set was partitioned into five equally-sized groups. In every run, one of the five groups was chosen as the test set, whereas the other four groups constituted the training sets. The average of cross-validation accuracy was calculated as where $k=5$

The use of both hold-out test and five-fold cross-validation procedures ensured that the evaluation of the classification accuracy, stability, and generalization capacity was made in detail.

V. RESULTS AND DISCUSSION

Hyperspectral Reflectance Signature Analysis

The averaged hyperspectral reflectance signatures of the sixteen soil samples obtained by the use of ASD FieldSpec Spectroradiometer are illustrated in Figure 4. For each soil sample, the mean value of ten repeated spectral measurements at each sampling point is shown in this figure.



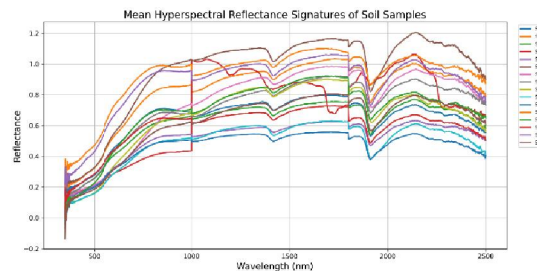


Fig. 4 Hyperspectral Reflectance Signatures of Soil Samples

It is clearly seen that the spectral signatures have different values in VIS, NIR, and SWIR. There are significant absorptions around 1400 nm and 1900 nm. These values are usually attributed to soil moisture and hydroxyl absorption. Also, some differences can be noticed in the 2200–2400 nm range and these are related to soil mineralogical differences. Therefore, it can be concluded that hyperspectral spectroscopy has managed to capture the specific reflectance characteristics of the soils that can be used for the spectral signature analysis.

Principal Component Analysis (PCA) Results

To decrease the number of features in the hyperspectral dataset, PCA was used to reduce the number of features from the initial 2051 spectral bands to 50 components.

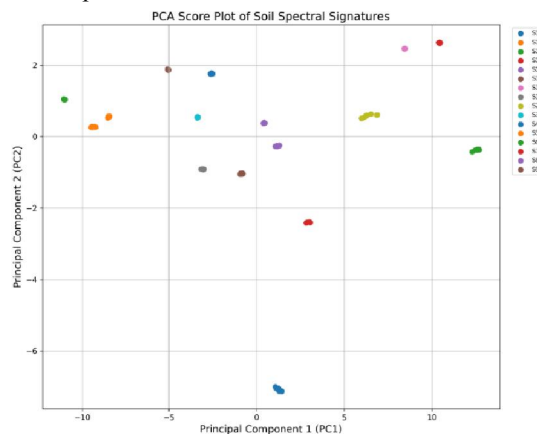


Fig. 5 PCA Score Plot of Soil Spectral Signatures.

It is evident from the PCA score plot (Figure 5) that the distribution of the soil samples in the new reduced feature space proves that the selected components still carry most of the information about the significant spectral differences between the sixteen sampling locations.

Classification Performance of the Random Forest Model

The Random Forest (RF) classifier was trained based on the hyperspectral dataset reduced by the PCA algorithm with 50 principal components. To assess the model performance, an 80:20 stratified train-test split was used in order to preserve equal proportions of all sixteen classes. In this work, the Random Forest algorithm was implemented with 300 decision trees and a fixed random state set to 42. The classification accuracy of the RF classifier reached 100% for the testing dataset. Besides, precision, recall, and F1-score of all classes were equal to 1.00, which proves the capability of the model to identify accurately the hyperspectral reflectance spectra of all sixteen soil samples. These results demonstrate that the spectral characteristics extracted from the ASD FieldSpec measurements can be used for successful soil classification performed in controlled laboratory settings. The confusion matrix presented in Fig. 6. Confirms the classification performance of the Random Forest model.



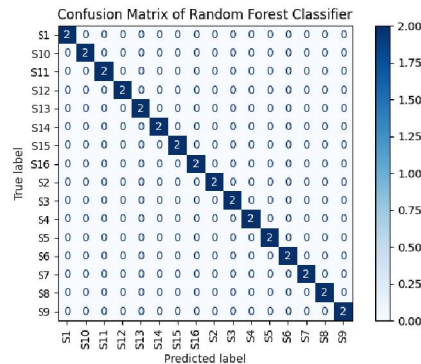


Fig. 6 Confusion Matrix of Random Forest

The confusion matrix presented in Fig. 6. confirms the classification performance of the Random Forest model. All samples of the test set belong to the correct classes, which results in non-zero matrix elements only on the main diagonal. This implies that there was no case of any misclassification during testing. To assess the stability of the classifier, five-fold cross-validation was done. The Random Forest classifier attained five-fold cross-validation accuracies of 1.00 for all folds, with a mean cross-validation accuracy of 1.00 and a standard deviation of 0.00.

Support Vector Machine Classification Results

Support Vector Machine (SVM) classifier was developed using the transformed PCA-based hyperspectral dataset under similar conditions used in the training of Random Forest classifier. The support vector machine model adopted Radial Basis Function (RBF) kernel with default hyperparameters to classify the sixteen soil classes. An 80:20 stratified train-test split was adopted for testing purposes. The SVM classifier resulted in overall classification accuracy of 100% for the testing data set. Besides, the precision, recall and F1 score for all soil classes were 1.00.

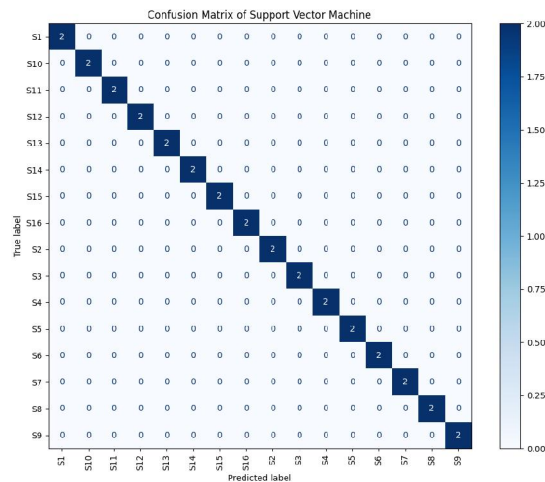


Fig. 7 Confusion Matrix of Support Vector Machine

Fig. 7 shows the confusion matrix where all testing samples have been perfectly classified to the corresponding soil class categories with no off-diagonal misclassification. In order to examine the robustness of the classifier, five-fold cross-validation was conducted. The accuracy of the SVM classifier in five-fold cross-validation was 1.00, which provided the mean accuracy of 1.00 and standard deviation of 0.00 for the SVM model.



Comparative Performance Analysis of Random Forest and SVM

In order to compare the performance of the applied machine learning algorithms, classification results of the Random Forest and SVM classifiers based on the same hyperspectral dataset transformed with PCA and identical experimental conditions were compared. Both classifiers used an 80/20 stratified train-test split and were validated with five-fold cross-validation.

According to the experimental results, Random Forest and Support Vector Machine classifiers were able to achieve 100% accuracy on the testing dataset. At the same time, the precision, recall, and F1-score of both classifiers were equal to 1.00. It means that all sixteen soil sampling locations could be correctly classified without any errors. Moreover, both models obtained 1.00 of mean accuracy with standard deviation of 0.00 in five-fold cross-validation. However, although the two classifiers exhibited similar performance measures, their learning processes are different. While the Random Forest classifier is an ensemble learning model that combines the outputs of several decision trees to form a more robust classifier tolerant to noise, the Support Vector Machine is very effective because it finds the best hyperplane that maximizes the margins between classes. It is worth noting that the same performance of the two classifiers was due to the laboratory environment in which the measurements were carried out, as well as due to the collection of several repeated scans of the hyperspectral data for each sample. This way of collecting the data has led to highly discriminative spectra that made it possible for both machine learning models to perfectly classify the sixteen soil classes. Thus, the obtained results show the efficiency of the use of the hyperspectral reflectance spectra in the case of classification of soils under controlled laboratory conditions.

TABLE II: Performance Comparison of Machine Learning Models

Performance Metric	Random Forest	Support Vector Machine
Accuracy	100%	100%
Precision	1.00	1.00
Recall	1.00	1.00
F1 - score	1.00	1.00
Mean Cross-Validation Accuracy	1.00	1.00
Standard Deviation	0.00	0.00

According to the experimental results, Random Forest and Support Vector Machine classifiers were able to achieve 100% accuracy on the testing dataset. At the same time, the precision, recall, and F1-score of both classifiers were equal to 1.00. It means that all sixteen soil sampling locations could be correctly classified without any errors. Moreover, both models obtained 1.00 of mean accuracy with standard deviation of 0.00 in five-fold cross-validation. However, although the two classifiers exhibited similar performance measures, their learning processes are different. While the Random Forest classifier is an ensemble learning model that combines the outputs of several decision trees to form a more robust classifier tolerant to noise, the Support Vector Machine is very effective because it finds the best hyperplane that maximizes the margins between classes. It is worth noting that the same performance of the two classifiers was due to the laboratory environment in which the measurements were carried out, as well as due to the collection of several repeated scans of the hyperspectral data for each sample. This way of collecting the data has led to highly discriminative spectra that made it possible for both machine learning models to perfectly classify the sixteen soil classes. Thus, the obtained results show the efficiency of the use of the hyperspectral reflectance spectra in the case of classification of soils under controlled laboratory conditions.

Five-Fold Cross-Validation Results

In order to estimate the robustness of the created models of machine learning, five-fold cross-validation was done on the entire PCA-transformation hyperspectral dataset. Each dataset was divided into five subsets randomly. One out of the five subsets was used as a validation set in each iteration, whereas the other four subsets were utilized for the



training purpose. Such process was iterated five times, and it made sure that each of the five subsets acted as a validation set at least once.

It can be observed from the cross-validation results that both RF and SVM classifiers had obtained 100% classification accuracy for all five iterations. The mean cross-validation accuracy of both the classifiers was 1.00, with zero standard deviation in this respect.

TABLE III: Five-Fold Cross-Validation Results of Random Forest and Support Vector Machine

Cross-Validation Fold	Random Forest	Support Vector Machine
Fold 1	1.00	1.00
Fold 2	1.00	1.00
Fold 3	1.00	1.00
Fold 4	1.00	1.00
Fold 5	1.00	1.00
Mean Accuracy	1.00	1.00
Standard Deviation	0.00	0.00

The consistent cross-validation performance reveals that the use of PCA-transformed hyperspectral features has resulted in the retention of discriminating spectral data required for the classification of soil samples. In addition to it, consistent cross-validation performance across all five iterations shows the robustness of both the machine learning algorithms under laboratory settings of the present study. Yet, future studies may use field datasets to test the generalization ability of the developed classifiers.

VI. DISCUSSION

These results indicate that the spectral signatures obtained using the ASD FieldSpec Spectroradiometer carry enough spectral information for accurate soil classification. There was variation in the reflectance spectra across the VIS, NIR, and SWIR regions, which is an indication that there were differences in the physical and chemical characteristics of the soil samples. In the dimensionality reduction process using Principal Component Analysis (PCA), the number of bands in the hyperspectral data set was reduced from 2151 bands to 50 PCs. This helped to reduce computational costs while retaining the most important spectral information.

The accuracy of the two machine learning classifiers used in this study was 100%, and the precision, recall, and F1 score were 1.00. This excellent result was due to the laboratory controlled conditions, proper calibration with the white reference, preparation of samples, and spectral measurement of the soils samples repeatedly. This resulted in a very consistent spectral signature with good separability among the sixteen soils classes. Even though the two classifiers performed equally well, their classification processes were different. The Random Forest algorithm involves an ensemble of decision trees while the Support Vector Machine algorithm involves the determination of the optimal separating hyperplane. It can be seen that the results obtained using the two algorithms are quite similar. This is due to the fact that the PCA-transformed hyperspectral variables contained enough information for discriminating soil types regardless of which classifier was used.

One point that needs to be considered is that the 100% accuracy achieved does not necessarily reflect the performance of the algorithm when applied to independent soil field datasets. Instead, it reflects the repeatability of the information contained within the laboratory acquired data set.

VII. CONCLUSION

In this study, a machine learning framework for classification of soil hyperspectral reflectance spectra measured by ASD FieldSpec spectroradiometer was designed. After the preprocessing of hyperspectral measurements taken in the 350-2500 nm wavelength range, PCA technique was applied for dimensionality reduction from 2151 bands down to 50 principal components. Further, the resulted feature set was classified by means of RF and SVM algorithms. The



experimental results showed that both classifiers obtained 100% classification accuracy with precision, recall and F1-score equal to 1.00. Moreover, the five-fold cross-validation technique proved that the developed models are stable and consistent, since the average accuracy is 1.00 and its standard deviation equals to 0.00. Thus, it is possible to conclude that hyperspectral reflectance spectra contain very discriminative features for soil classification under laboratory conditions.

In general, the study highlights the potentialities of using hyperspectral spectroscopy and machine learning for rapid and non-destructive soil characterization. The described approach may be a basis for further researches on the topic of developing a reliable soil classification system for precision agriculture and environmental monitoring purposes.

K. M. De Oliveira, J. V. F. Gonçalves, R. Falcioni, C. A. De Oliveira, D. F. da Silva Haubert, W. A. Mendonça, L. G. T. Crusiol, R. B. De Oliveira, A. S. Reis, E. César, and M. Nanni, "Classification of soil horizons based on VisNIR and SWIR hyperspectral images and machine learning models," *Remote Sens. Appl.: Soc. Environ.*, Art. no. 101362, 2024, doi: 10.1016/j.rsase.2024.101362.

REFERENCES

1. G. D. Chate and S. S. Bhamare, "Machine learning approaches for soil image classification: A systematic review," *Int. J. Innov. Sci. Res. Technol.*, vol. 10, no. 4, pp. 825–829, 2025, doi: 10.38124/IJISRT/25APR045.
2. H. S. Prasantha, M. M. Chatterjee, P. S. Roshitha, and V. Roshitha, "Soft computing approaches for hyperspectral image classification," *ICTACT J. Soft Comput.*, vol. 10, no. 4, pp. 2034–2040, 2020, doi: 10.21917/ijsc.2020.0304.
3. H. Yu, B. Kong, Q. Wang, X. Liu, and X. Liu, "Hyperspectral remote sensing applications in soil: A review," in *Hyperspectral Remote Sensing: Theory and Applications*. Academic Press, 2020, pp. 269–291, doi: 10.1016/B978-0-08-102894-0.00011-5.
4. H. Petersson, D. Gustafsson, and D. Bergström, "Hyperspectral image analysis using deep learning—A review," in *Proc. 6th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, 2016, doi: 10.1109/IPTA.2016.7820963.
5. T. Mahmood, M. Usman, and C. Conrad, "Selecting relevant features for random forest-based crop type classifications by spatial assessments of backward feature reduction," *PFG—J. Photogramm., Remote Sens. Geoinf. Sci.*, vol. 93, no. 2, pp. 173–196, 2025, doi: 10.1007/s41064-024-00329-4.
6. C. Piccini, K. Metzger, G. Debaene, B. Stenberg, S. Götzinger, L. Borůvka, T. Sandén, L. Bragazza, and F. Liebisch, "In-field soil spectroscopy in Vis–NIR range for fast and reliable soil analysis: A review," *Eur. J. Soil Sci.*, vol. 75, no. 2, Art. no. e13481, 2024, doi: 10.1111/ejss.13481.
7. F. Canero, V. Rodriguez-Galiano, and D. Aragonés, "Machine learning and feature selection for soil spectroscopy: An evaluation of random forest wrappers to predict soil organic matter, clay, and carbonates," *Heliyon*, vol. 10, Art. no. e30228, 2024, doi: 10.1016/j.heliyon.2024.e30228.
8. K. M. De Oliveira, R. Falcioni, J. V. F. Gonçalves, C. A. De Oliveira, W. A. Mendonça, L. G. Crusiol, R. B. De Oliveira, R. H. Furlanetto, A. S. Reis, and M. Nanni, "Rapid determination of soil horizons and suborders based on VIS-NIR-SWIR spectroscopy and machine learning models," *Remote Sens.*, vol. 15, no. 19, Art. no. 4859, 2023, doi: 10.3390/rs15194859.
9. K. M. De Oliveira, J. V. F. Gonçalves, R. Falcioni, C. A. De Oliveira, D. F. da Silva Haubert, W. A. Mendonça, L. G. T. Crusiol, R. B. De Oliveira, A. S. Reis, E. César, and M. Nanni, "Classification of soil horizons based on VisNIR and SWIR hyperspectral images and machine learning models," *Remote Sens. Appl.: Soc. Environ.*, Art. no. 101362, 2024, doi: 10.1016/j.rsase.2024.101362.
10. K. M. De Oliveira, J. V. F. Gonçalves, R. H. Furlanetto, C. A. De Oliveira, W. A. Mendonça, D. F. da Silva Haubert, L. G. Crusiol, R. Falcioni, R. B. De Oliveira, A. S. Reis, A. E. A. Ecker, and M. Nanni, "Predicting particle size and soil organic carbon of soil profiles using VIS-NIR-SWIR hyperspectral imaging and machine learning models," *Remote Sens.*, vol. 16, Art. no. 2869, 2024, doi: 10.3390/rs16162869.



11. Gholizadeh, L. Borůvka, M. Saberioon, and R. Vašát, "A memory-based learning approach as compared to other data mining algorithms for the prediction of soil texture using diffuse reflectance spectra," *Remote Sens.*, vol. 8, no. 4, Art. no. 341, 2016, doi: 10.3390/rs8040341.
12. Ghosh, T. Mitran, and P. K. Mani, "Estimation of soil nitrogen content influenced by different nitrogen-based management practices within rice-based cropping using diffuse reflectance spectroscopy and machine learning," *Commun. Soil Sci. Plant Anal.*, vol. 56, pp. 856–878, 2025, doi: 10.1080/00103624.2024.2433703.
13. T. Hu, C.-C. Qi, M. Wu, T. Rennert, Q.-S. Chen, L. Chai, and Z. Lin, "Classification of arsenic contamination in soil across the EU by Vis-NIR spectroscopy and machine learning," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 134, Art. no. 104158, 2024, doi: 10.1016/j.jag.2024.104158.
14. K. Karyotis, N. Tsakiridis, N. V. Tziolas, N. Samarinas, E. Kalopesa, P. Chatzimisios, and G. Zalidis, "On-site soil monitoring using photonics-based sensors and historical soil spectral libraries," *Remote Sens.*, vol. 15, no. 6, Art. no. 1624, 2023, doi: 10.3390/rs15061624.
15. V. Khosravi, F. D. Ardejani, S. Yousefi, and A. Aryafar, "Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods," *Geoderma*, vol. 318, pp. 29–41, 2018, doi: 10.1016/j.geoderma.2017.12.025.
16. J. Padarian, B. Minasny, and A. B. McBratney, "Using deep learning to predict soil properties from regional spectral data," *Geoderma Regional*, vol. 16, Art. no. e00198, 2019, doi: 10.1016/j.geodrs.2018.e00198.
17. Pan, S. Cai, M. Zhao, H. Cheng, H. Yu, S. Du, J. Du, and F. Xie, "Predicting the surface soil texture of cultivated land via hyperspectral remote sensing and machine learning: A case study in Jianghuai hilly area," *Appl. Sci.*, vol. 13, no. 16, Art. no. 9321, 2023, doi: 10.3390/app13169321.
18. T. Shi, H. Liu, Y. Chen, T. Fei, J. Wang, and G. Wu, "Spectroscopic diagnosis of arsenic contamination in agricultural soils," *Sensors*, vol. 17, no. 5, Art. no. 1036, 2017, doi: 10.3390/s17051036.
19. Vibhute and K. Kale, "Mapping several soil types using hyperspectral datasets and advanced machine learning methods," *Results Opt.*, vol. 12, Art. no. 100503, 2023, doi: 10.1016/j.rso.2023.100503.
20. Vibhute, K. Kale, and S. Gaikwad, "Machine learning-enabled soil classification for precision agriculture: A study on spectral analysis and soil property determination," *Appl. Geomatics*, 2024, doi: 10.1007/s12518-023-00546-3

