Iraqi National Journal of Earth Science
www.earth.mosuljournals.com

Using Artificial Intelligence and Deep Learning Algorithms to Extract Land Features from High Resolution Pleiades Data.

Saad Mahmood Sulaiman [1*] , **Alyaa Abbas Ali Al-Attar** [2] , **Mustafa Ridha Mezaal** [3]
[1,2,3] *Northern Technical University*

Article information

Received: 09- Apr -2024

Revised: 28- May -2024

Accepted: 17- Jul -2024

Available online: 01- Jul – 2025

Keywords:
Soil Texture Prediction
Machine Learning
Grid Search
Random Search
Optimization

Correspondence:
Name: Saad Mahmood Sulaiman
Email: saad.mahmoodgs@ntu.edu.iq

ABSTRACT

This work looks at the use of machine learning algorithms to predict soil composition (clay, gravel, sand, and silt) using remotely sensed data, providing a cost-effective and scalable solution for large-area soil mapping. This paper aims to optimize and compare grid search and random search for improving the performance of the machine learning models for soil texture prediction in the study area using Sentinel-1A SAR data and ASTER Global Digital Elevation Model (GDEM), and topographic information. Five machine learning models—Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), Decision Tree (DT), and Multilayer Perceptron Regressor (MLP Regressor)—are examined. Grid search and random search approaches are used to optimize hyperparameters and improve model performance. After hyperparameter adjustment using grid and random searches, the DT model achieved near-perfect accuracy (RMSE $\leq$ 0.029, MAE $<$ 0.021, R2 = 1.000). The MLP Regressor model also performed well in random search optimization (RMSE = 0.038, MAE = 0.03, R2 = 1.000), outperforming grid search. Based on the presented results, the Decision Tree model appears to be the most suitable choice for predicting (clay, silt, and sand) soil composition, and Multilayer Perceptron Regressor (MLP Regressor) is the most suitable choice for predicting gravel composition. The improved models may be used in large-scale soil mapping projects, allowing for more informed decisions in agriculture, environmental management, and land use planning.

DOI: 10.33899/earth.2024.148645.1264, ©Authors, 2025, College of Science, University of Mosul.
This is an open access article under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/) .

استخدام الذكاء الاصطناعي وخوارزميات التعلم العميق لاستخراج معالم الأرض من بيانات بلياد عالية الدقة

سعد محمود سليمان [*1]، علياء عباس علي العطار [2]، مصطفى رضا مزعل [3]

الجامعة التقنية الشمالية[1،2،3]

معلومات الارشفة

تاريخ الاستلام: 09− ابريل −2024

تاريخ المراجعة: 28− مايو −2024

تاريخ القبـول: 17− يوليو −2024

تاريخ النشر الالكتروني: 01− يوليو −2025

الكلمات المفتاحية:

التنبؤ بنسيج التربة

التعلم الالي

بحث الشبكة

بحث عشوائي

تحسين

المراسلة:

الاسم: سعد محمود سليمان

Email: saad.mahmoodgs@ntu.edu.iq

الملخص

يتناول هذا البحث استخدام خوارزميات التعلم الآلي للتنبؤ بتركيبة التربة (الطين والحصى والرمل والطمي) باستخدام بيانات الاستشعار عن بعد، مما يوفر حلاً فعالاً من حيث التكلفة وقابل للتطوير لرسم خرائط التربة في مناطق كبيرة. الهدف من هذا البحث هو تحسين ومقارنة البحث الشبكي والبحث العشوائي لتحسين أداء نماذج التعلم الآلي للتنبؤ بنسيج التربة في منطقة الدراسة باستخدام بيانات رادار SAR Sentinel-1A ونموذج الارتفاع الرقمي العالمي ASTER (GDEM) والمعلومات الطبوغرافية. لتدريب النموذج والتحقق من صحته. يتم فحص خمسة نماذج للتعلم الآلي – الانحدار الخطي (LR)، وانحدار متجه الدعم (SVR)، والغابة العشوائية (RF)، وشجرة القرار (DT)، ومتغير الانحدار متعدد الطبقات (MLP Regressor). يتم استخدام البحث الشبكي وأساليب البحث العشوائي لتحسين المعلمات الفائقة وتحسين أداء النموذج. بعد تعديل المعلمات الفائقة باستخدام البحث الشبكي والعشوائي، حقق نموذج DT دقة شبه مثالية (RMSE ≤ 0.029، MAE < 0.021، R2 = 1.000). كما أظهر نموذج MLP Regressor أداءً جيدًا في تحسين البحث العشوائي (RMSE = 0.038، MAE = 0.03، R2 = 1.000)، متفوقًا على البحث الشبكي. بناءً على النتائج المقدمة، يبدو أن نموذج شجرة القرار هو الخيار الأكثر ملاءمة للتنبؤ بتكوين التربة (الطين والطمي والرمل) ونموذج MLP Regressor هو الخيار الأكثر ملاءمة للتنبؤ بتكوين الحصى. يمكن استخدام النماذج المحسنة في مشاريع رسم خرائط التربة واسعة النطاق، مما يسمح باتخاذ قرارات أكثر استنارة في الزراعة والإدارة البيئية وتخطيط استخدام الأراضي.

DOI: 10.33899/earth.2024.148645.1264, ©Authors, 2025, College of Science, University of Mosul.
This is an open access article under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/) .

Introduction

Soil texture prediction and mapping are critical in many agricultural, environmental, and land management applications (Laborczi et al., 2015; Khalil et al., 2016). Soil texture mapping has been widely investigated utilizing many ways, among them the remote sensing, laboratory spectra, and field-based technologies (Gomez et al., 2019). Several researchers have looked into how machine learning models may be integrated with various data sources to predict soil texture accurately. Several studies have shown that machine learning approaches and radar data work well for mapping soil texture. Forkuor et al. (2017) showed that random forest regression can improve indigenous soil knowledge in West Africa at low cost and effort. Ana et al. (2022) observed that random forest provided the most accurate forecasts for clay, silt, and sand concentrations in the Amazon area, especially when the P-band of airborne radar was included as a covariate. Similarly, Rengma et al. (2023) employed a random forest regression (RFR) model to map soil texture and organic carbon in the mid-Himalayas, producing very accurate results.

Several researchers have used synthetic aperture radar (SAR) data for soil texture mapping. Periasamy et al. (2019) used synthetic aperture radar to successfully identify the sandy loam (23%) and clay (35%) soil texture classes. Niang et al. (2014) discovered that using RADARSAT-2 polarimetric SAR data as covariates considerably enhanced the accuracy of digital mapping for soil surface texture when compared to regular kriging. Bousbih et al. (2019) used radar and optical data from Sentinel-1 and Sentinel-2 to analyze soil texture in Tunisia, reaching a 65% overall accuracy with the random forest (RF) method. Their findings showed that the soil moisture indicator derived from combined Sentinel-1 and Sentinel-2 data produced the best classifications, implying that combining these datasets with soil moisture indicators can improve mapping accuracy and soil texture estimation (Bousbih et al. 2019).

Grid search and random search are two prominent optimization methods for hyperparameter tweaking in machine learning. In the domain of remote sensing and soil texture prediction, these strategies have been used to improve machine learning model performance. Grid search includes searching exhaustively through a preset set of hyperparameter values, whereas random search selects hyperparameter values randomly from a specified range. Both strategies have demonstrated potential in optimizing hyperparameters for machine learning models in remote sensing applications (Yasser et al., 2023; Vladyslav et al., 2022). Grid search is a methodical method for covering the full search space, although it can be computationally costly. Random search, on the other hand, is more computationally efficient since it only investigates a limited part of the search universe. Random search is more efficient for hyperparameter optimization than grid search because it finds better models with a less computational budget and covers a broader configuration space (Bergstra and Bengio, 2012).

The purpose of this work is to evaluate grid and random search optimization strategies for fine-tuning different machine learning algorithms for soil texture prediction using Sentinel-1A SAR data and topographic information. The paper aims to analyze two optimization strategies for models including linear regression (LR), linear support vector machine (SVM), decision tree (DT), random forest (RF), and artificial neural networks (ANN).

Study area

The current study focuses on a region in the eastern part of the Nineveh Governorate that stretches approximately 30 kilometers along the Great Zab River, a significant tributary of the Tigris River, from Kalak area to Al-Gwair area (Fig. 1). Geologically, the area is a part of northern Iraqi Folded Zone, which has a complicated structural and tectonic environment.

Materials and Methods

Figure 2 shows several data sources used in the process, including remote sensing data (Sentinel-1A SAR pictures and ASTER GDEM) and field data (soil samples). The data sources are used to calculate backscattering coefficients, relief characteristics, and field soil composition data, and Table 1 shows the input dataset.

Table 1: Input dataset.

Sentinel-1A SAR data	VV, VH Polarizations
ASTER Global Digital Elevation Model	(GDEM)
topographic information	Gravel, Sand, Silt, and Clay content

The fundamental fact units utilized in this examine are the ASTER GDEM and the Sentinel-1A Synthetic Aperture Radar (SAR) image. The 2014 introduction of Sentinel-1A SAR information has made some of makes applications of, including monitoring modifications in land cover, agriculture, forestry, and catastrophe management. Depending on the mode and polarization, Sentinel-1A's SAR imaging has a spatial resolution of five to forty meters, which enables an in-depth analysis of the features on Earth's surface. Unlike optical sensors, SAR operates within the microwave vicinity of the electromagnetic spectrum, allowing it to function day or nighttime and through clouds. Satellite records from Sentinel-1A (S-1A) include the C-band dual-polarization channels (VV and VH) with a 12-day repeating cycle. Two Sentinel-1A photos have been received for this research on January 20, 2024.

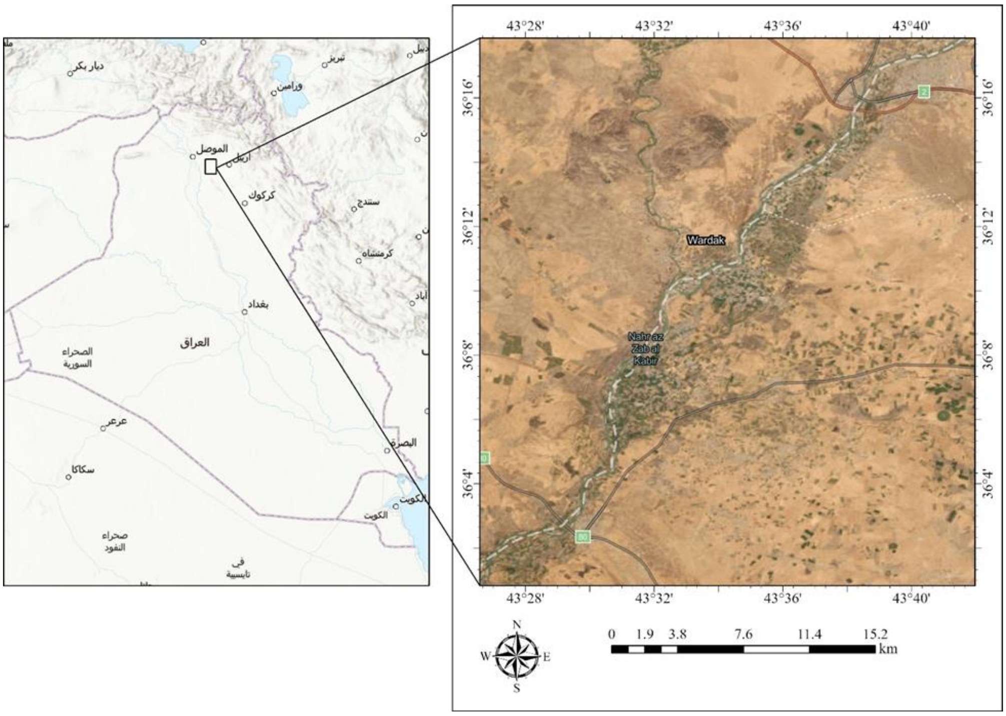

Fig. 1. Map of the study area.

At 75 different places within the research region, which extends from (36° 16′ 02″ N, 43° 38′ 27″ E) and ends at (36° 02′ 27″ N, 43° 29′ 05″ E), soil samples are taken. Sample locations are chosen along the Greater Zap River at about equal intervals of 400 m. A 500-gram soil sample is taken from the subsurface of each location (50 cm deep). The samples are forwarded to the lab for analysis. To test the qualities of the soil, the water content, sieve analysis, and hydrometer analysis are carried out.

NASA and the Japanese Ministry of Economy, Trade, and Industry (METI) collaborated to develop ASTER, the Advanced Spaceborne Thermal Emission and Reflection Radiometer. NASA's Terra probe collected high-resolution topography data for the Earth's surface using stereo-pair photos processed by the ASTER sensor. With a spatial resolution of around 30 m, the ASTER GDEM provides global coverage and is suitable for a variety of applications such as landform mapping, topography analysis, and natural resource management. Its elevation data is critical to many sectors.

These data sources are merged to form a geodatabase, which serves as the foundation for the modeling process. The data is then separated into three subsets: train (60%), validation (20%), and test (20%). Before proceeding, Table 2 shows the VIF values for each variable in the dataset after doing a multicollinearity analysis. Six variables, including the Topographic Wetness Index (TWI), Slope, LS Factor, Landform, Flow Accumulation, and Catchment Area, were removed because their VIF values exceeded 10. This shows that there is a substantial problem with collinearity with these variables, which may cause difficulties in understanding the model's regression findings. The VIF values for the other variables are modest. VH, VV, Vertical Distance to Channel Network, Valley Depth, Slope Height, Mid Slope Position, Melton Ruggedness Number, Convexity, Convergence Index, and Elevation were among the variables retained when the VIF was less than 10. These low VIF ratings indicate fewer major problems with collinearity. Eliminating highly collinear variables reduces the dataset's redundancy and makes it simpler to identify how the remaining variables in the regression model impact each other individually. Although removing collinear variables improves interpretability, it is important to note that doing so may result in the loss of some data.

Table 2: Accuracy without extraction for classification algorithms.

Variable	VIF
VH	2.504
VV	2.622
Vertical Distance to Channel Network	4.058
Valley Depth	5.858
Topographic Wetness Index	133.526
Slope	50.256
Slope Height	4.542
Mid Slope Position	3.021
Melton Ruggedness Number	4.559
LS Factor	37.039
Landform	48.203
Flow Accumulation	17.14
Convexity	4.145
Convergence Index	4.263
Catchment Area	58.546
Elevation	1.551

The modeling method makes use of machine learning techniques, including linear regression (LR), linear Support Vector Machines (SVM), Radial Basis Function (RBF) SVM, Decision Trees (DT), Random Forests (RF), and Artificial Neural Networks (ANN). These algorithms are trained on the training set, optimized on the validation set, and then assessed on the test set. The models are optimized using two basic strategies: grid search and random search. These optimization procedures are used to fine-tune hyperparameters in machine learning algorithms, which have a significant impact on prediction performance.

The grid search technique defines a predetermined set of hyperparameter values, and the model is trained and tested for every possible combination of these hyperparameters. This method is comprehensive and guarantees that the optimal combination of hyperparameters within the specified range is found. However, it can be computationally expensive, particularly if there are a large number of hyperparameters with various values. Rather than doing an exhaustive search, the random search approach chooses hyperparameter values at random from a predetermined distribution (such as uniform, normal, or log-normal). This approach may be more efficient than grid search, especially in high-dimensional spaces, since it searches the hyperparameter space more thoroughly and has a better probability of finding the optimal combination of hyperparameters with fewer iterations.

The best-performing model is chosen based on the evaluation metrics produced from the validation set after the models have been improved using different search techniques. The chosen model is then utilized to generate projected soil composition maps (clay, gravel, sand, and silt) for the research region.

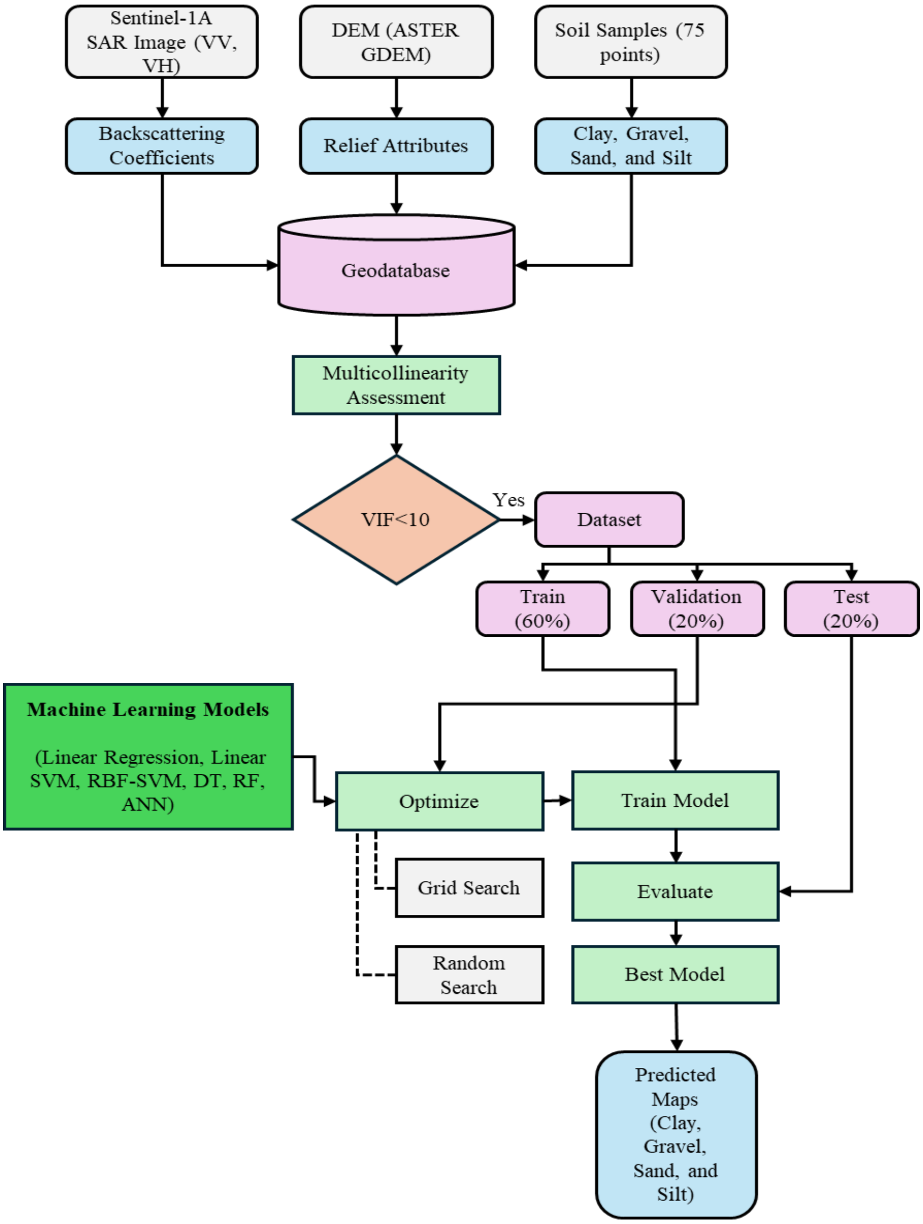
Sentinel-1A SAR Image (VV, VH)
DEM (ASTER GDEM)
Soil Samples (75 points)
Backscattering Coefficients
Relief Attributes
Clay, Gravel, Sand, and Silt
Geodatabase
Multicollinearity Assessment
VIF<10
Yes
Dataset
Train (60%)
Validation (20%)
Test (20%)
Machine Learning Models
(Linear Regression, Linear SVM, RBF-SVM, DT, RF, ANN)
Optimize
Train Model
Grid Search
Evaluate
Random Search
Best Model
Predicted Maps (Clay, Gravel, Sand, and Silt)

Fig. 2. Flowchart of the proposed methodology for soil composition prediction using machine learning with optimized hyperparameters.

Results and Discussion

Table 3 compares the performance metrics (RMSE, MAE, and R2) of various machine learning models for predicting soil compositions (clay, gravel, sand, and silt). The models tested include SVR, RF, DT, LR, and MLP Regressor. The data compares the performance of these models with their default hyperparameters to their performance after improving the hyperparameters using grid search and random search methods.

Table 3: Accuracy with extraction for classification algorithms.

Target Variable	Model	RMSE			MAE			R2		
		Default	Grid	Random	Default	Grid	Random	Default	Grid	Random
Clay	SVR	0.072	0.068	0.067	0.062	0.061	0.058	0.900	0.896	0.900
	RF	0.094	0.022	0.008	0.057	0.009	0.003	0.826	0.989	0.999
	DT	0.095	0.003	0.002	0.057	0.002	0.002	0.825	1.000	1.000
	LR	0.169	0.155	0.162	0.143	0.119	0.124	0.441	0.467	0.485
	MLP	0.087	0.011	0.009	0.070	0.008	0.007	0.853	0.998	0.998
Gravel	SVR	14.717	2.535	2.535	10.119	0.809	0.809	0.188	0.978	0.978
	RF	10.802	1.523	0.772	8.400	0.581	0.311	0.562	0.992	0.998
	DT	10.736	0.512	0.895	8.089	0.353	0.186	0.568	1.000	1.000
	LR	13.505	13.893	13.893	10.841	10.841	10.841	0.316	0.330	0.330

	MLP	16.596	0.424	0.038	13.192	0.295	0.030	0.033	0.999	1.000
Sand	SVR	13.115	2.236	2.236	9.074	0.765	0.765	0.133	0.976	0.976
	RF	11.289	1.060	1.070	8.736	0.357	0.422	0.358	0.995	0.995
	DT	11.242	0.117	0.097	8.689	0.050	0.078	0.363	1.000	1.000
	LR	10.213	10.213	10.213	8.568	8.568	8.568	0.474	0.474	0.474
	MLP	19.941	0.249	0.249	16.198	0.185	0.190	0.999	1.000	1.000
Silt	SVR	1.364	0.018	0.018	0.062	0.059	0.059	0.900	0.996	0.996
	RF	1.738	0.032	0.038	0.057	0.013	0.018	0.826	0.988	0.983
	DT	1.789	0.023	0.012	0.057	0.029	0.015	0.825	0.994	1.000
	LR	2.307	0.169	0.169	0.143	0.179	0.168	0.441	0.461	0.461
	MLP	1.051	0.040	0.029	0.070	0.032	0.021	0.853	1.000	1.000

Clay Composition Prediction

Among the models with default hyperparameters, the SVR model exhibits the best performance, with an RMSE of 0.072, an MAE of 0.062, and an R2 of 0.900. The Linear Regression model, on the other hand, shows the poorest performance with an RMSE of 0.169, an MAE of 0.143, and an R2 of 0.441. Hyperparameter optimization, through both grid search and random search techniques, significantly improved the performance of several models.

Random Forest and Decision Tree models show the most substantial improvements, achieving near-perfect performance (RMSE ≈ 0, MAE ≈ 0, and R^2 ≈ 1) after optimization. The MLP Regressor also exhibits significant improvements with an RMSE of 0.011 (grid search) and 0.009 (random search), an MAE of 0.008 (grid search) and 0.007 (random search), and an R^2 of 0.998 for both optimization techniques. The SVR model shows relatively minor improvements, with slightly lower RMSE and MAE values after optimization. The Linear Regression model displays modest improvements, with slightly lower RMSE and MAE values, and a higher R2 after optimization.

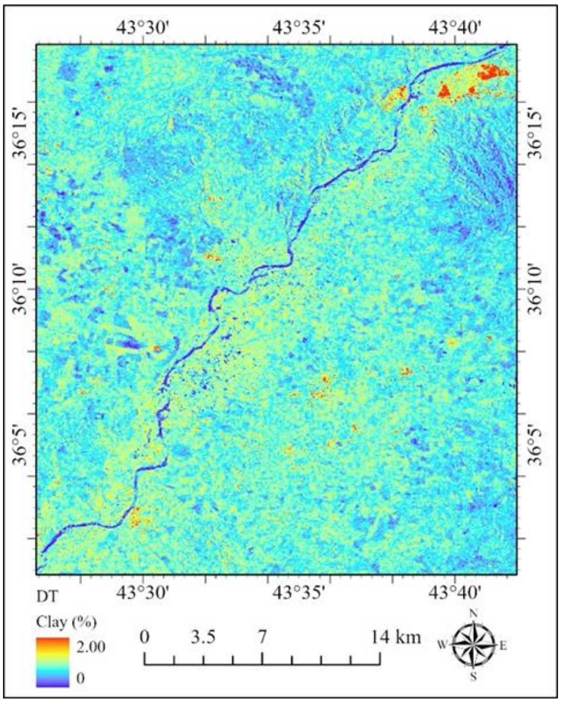

Fig. 3. Predicted clay composition with best machine learning models (DT).

Based on the presented results by Geographic Information System (GIS) software, the Decision Tree and Random Forest models appear to be the most suitable choices for predicting clay soil composition, exhibiting near-perfect performance after hyperparameter optimization. These models offer high accuracy while maintaining interpretability, as their decision-making process can be traced back through the tree structure. If computational efficiency is a concern, the MLP Regressor could be a viable alternative, as it achieved comparable performance to the tree-based models after optimization, while potentially being more computationally efficient for large datasets. Figure 3 presents the predicted soil maps for clay composition based on the best model.

Gravel Composition Prediction

Among the models with default hyperparameters, the Random Forest model exhibits the best performance, with an RMSE of 10.802, an MAE of 8.400, and an R^2 of 0.562. The MLP Regressor shows the poorest performance, with an RMSE of 16.596, an MAE of 13.192, and an R^2 of 0.033. Hyperparameter optimization, through both grid search and random search techniques, significantly improved the performance of several models. The MLP Regressor shows the most substantial improvements, with an RMSE of 0.424 (grid search) and 0.038 (random search), an MAE of 0.295 (grid search) and 0.030 (random search), and an R^2 of 0.999 (grid search) and 1.000 (random search). The Decision Tree model also exhibits significant improvements, achieving an RMSE of 0.512 (grid search) and 0.895 (random search), an MAE of 0.353 (grid search) and 0.186 (random search), and an R^2 of 1.000 for both optimization techniques. The Random Forest model shows improvements, with an RMSE of 1.523 (grid search) and 0.772 (random search), an MAE of 0.581 (grid search) and 0.311 (random search), and an R^2 of 0.992 (grid search) and 0.998 (random search). The SVR model exhibits substantial improvements, with an RMSE of 2.535 for both optimization techniques, an MAE of 0.809 for both optimization techniques, and an R^2 of 0.978 for both optimization techniques. The Linear Regression model shows minimal improvement, with slightly higher RMSE and R^2 values after optimization, but no change in MAE.

Based on the presented results by Geographic Information System (GIS) software, the MLP Regressor and Decision Tree models appear to be the most suitable choices for predicting gravel composition, exhibiting near-perfect performance after hyperparameter optimization using random search. These models offer high accuracy while maintaining computational efficiency, particularly for the Decision Tree model. The Random Forest model also demonstrated excellent performance after optimization, especially with random search, and could be considered as an alternative if ensemble models are preferred over individual tree models. Figure 4 presents the predicted soil maps for gravel composition based on the best model.

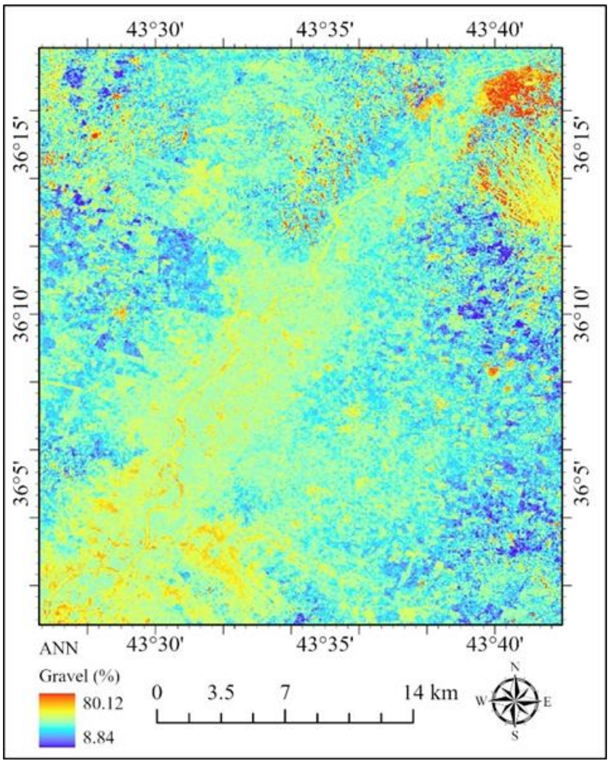

Fig. 4. Predicted gravel composition with best machine learning models (ANN).

Sand Composition Prediction

Among the models with default hyperparameters, the Linear Regression model exhibited the best performance, with an RMSE of 10.213, an MAE of 8.568, and an R^2 of 0.474. The MLP Regressor shows the poorest performance, with an RMSE of 19.941, an MAE of 16.198, and an R^2 of 0.999.

Hyperparameter optimization, through both grid search and random search techniques, significantly improves the performance of several models. The MLP Regressor shows the most substantial improvements, with an RMSE of 0.249 for both optimization techniques, an MAE of 0.185 (grid search) and 0.190 (random search), and an R^2 of 1.000 for both optimization techniques. The Decision Tree model also exhibits significant improvements, achieving an RMSE of 0.117 (grid search) and 0.097 (random search), an MAE of 0.050 (grid search) and 0.078 (random search), and an R^2 of 1.000 for both optimization techniques. The Random Forest model shows improvements, with an RMSE of 1.060 (grid search) and 1.070 (random search), an MAE of 0.357 (grid search) and 0.422 (random search), and an R^2 of 0.995 for both optimization techniques. The SVR model exhibits substantial improvements, with an RMSE of 2.236 for both optimization techniques, an MAE of 0.765 for both optimization techniques, and an R^2 of 0.976 for both optimization techniques. The Linear Regression model shows no improvement after hyperparameter optimization, as the performance metrics remained unchanged.

Based on the presented results by Geographic Information System (GIS) software, the Decision Tree and MLP Regressor models appear to be the most suitable choices for predicting sand composition, exhibiting near-perfect performance after hyperparameter optimization using both grid search and random search techniques. These models offer high accuracy while maintaining computational efficiency, particularly for the Decision Tree model. The Random Forest model also demonstrates excellent performance after optimization and could be considered as an alternative if ensemble models are preferred over individual tree models. Figure 5 presents the predicted soil maps for sand composition based on the best model.

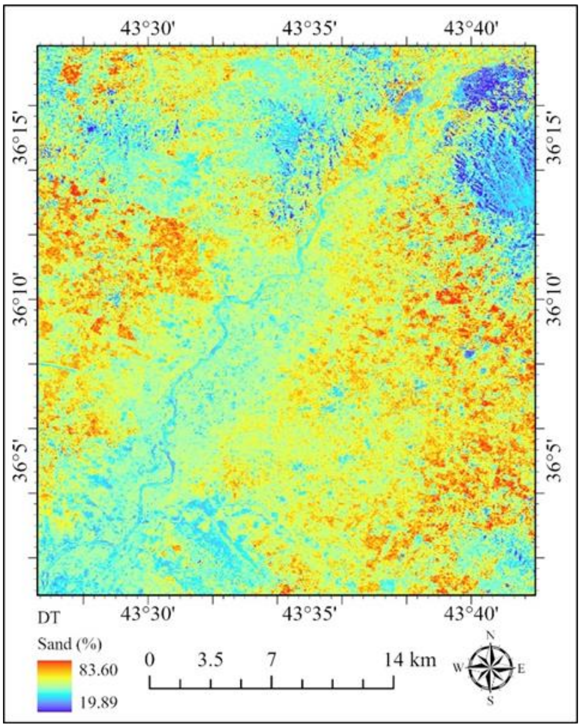

Fig. 5. Predicted sand composition with best machine learning models (DT).

Silt Composition Prediction

Among the models with default hyperparameters, the SVR model exhibits the best performance, with an RMSE of 0.072, an MAE of 0.062, and an R^2 of 0.900. The Linear Regression model shows the poorest performance, with an RMSE of 0.169, an MAE of 0.143, and an R^2 of 0.441. Hyperparameter optimization, through both grid search and random search techniques, significantly improves the performance of several models. The SVR model shows substantial improvements, with an RMSE of 0.018 for both optimization techniques, an MAE of 0.059 for both optimization techniques, and an R^2 of 0.996 for both optimization techniques. The MLP Regressor also exhibits significant improvements, achieving an RMSE of 0.040 (grid search) and 0.029 (random search), an MAE of 0.032 (grid search) and 0.021 (random search), and an R^2 of 1.000 for both optimization techniques. The Decision Tree model shows improvements, with an RMSE of 0.023 (grid search) and 0.012 (random search), an MAE of 0.029 (grid search) and 0.015 (random search), and an R^2 of 0.994 (grid search) and 1.000 (random search). The Random Forest model exhibits modest improvements, with an RMSE of 0.032 (grid search) and 0.038 (random search), an MAE of 0.013 (grid search) and 0.018 (random search), and an R^2 of 0.988 (grid search) and 0.983 (random search). The Linear Regression model shows no improvement after hyperparameter optimization, as the performance metrics remained unchanged.

Based on the presented results, the Decision Tree and MLP Regressor models appear to be the most suitable choices for predicting silt composition, exhibiting near-perfect performance after hyperparameter optimization using both grid search and random search techniques. These models offer high accuracy while maintaining computational efficiency, particularly for the SVR model. The Random Forest model also demonstrates excellent performance after optimization, especially with random search, and could be considered as an alternative if interpretability and computational efficiency are prioritized over ensemble models. Figure 6 presents the predicted soil maps for silt composition based on the best model.

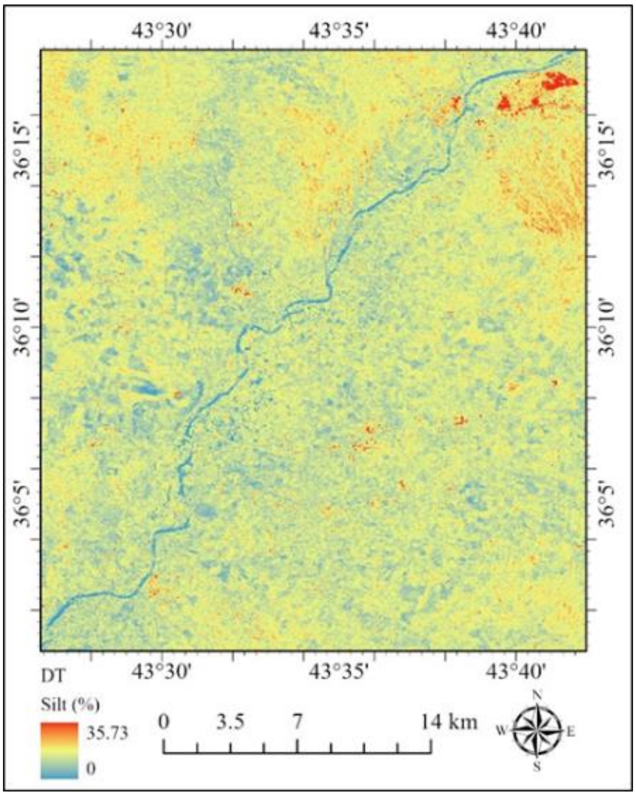

Fig. 6. Predicted silt composition with best machine learning models (DT).

The results of the above study are a progression on the trend whereby models that are complex, e.g., Decision Tree and Random Forest, give better results than simpler models such as Linear Regression in the case of soil composition prediction. They do it due to the tendency of complicated models to represent the nonlinear relationships and to detect the small but important details in the soil's structure that are probably used in the classification and prediction tasks.

In general, most models demonstrated more than a 5% increase in performance as a result of the parameter adjustments or grid search approach. Thus, this capability is an effective means of planning the hyperparameters of models for the better performance of prediction tasks, as they may be problem-specific. Both the grid search and the random search optimization strategies enhanced performance. Of the two studied, the random search is more successful in several diagnoses, such as the MLP Regressor and the Decision Tree. By suggesting that random search could be a better way of sampling through the hyperparameter space and also in the process come up with the best settings for these models, the soil composition prediction could be an easier task. While all methods remained just above the surface, in most rounds, the SVR can distinguish itself from the field by continuously attributing nearly perfect values to performance measures after the tuning of hyperparameters. This shows that SVR is one of the powerful models, which accurately predicts soil composition by sensing complicated interactions while preserving the wave of computational efficiency. Ensemble models like Random Forest prove strong and most successful after optimizing the hyperparameters. Which models are built by the accumulation of joint predictions of multiple decision trees rather than the simple addition of such predictions, and are capable of taking into account the nonlinearity and interactions in the data of soil compositions.

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

While complicated models such as MLP Regressor and Decision Tree achieved great accuracy, simpler models like Random Forest and SVR may be selected if interpretability and computing economy are important considerations. Decision trees, in particular, provide a simple and understandable framework that might be useful in determining the underlying correlations between soil attributes and composition.

Based on the presented results, the Decision Tree, MLP Regressor, and Random Forest models are the best choices for predicting soil composition, given their high accuracy, computational efficiency, and potential interpretability (in the case of Decision Trees). If ensemble-based forecasts are preferable, ensemble models such as Random Forest can be used, albeit interpretability may suffer.

Conclusion

In this work, we compared the efficacy of different machine learning models in predicting the composition of clay, gravel, sand, and silt in soil samples. The investigation includes evaluating the models' performance with default hyperparameters and optimized hyperparameters derived via grid search and random search approaches. The findings highlight the necessity of hyperparameter tweaking and model selection for accurate soil composition prediction. Among the models tested, the Multilayer Perceptron Regressor (MLP Regressor), Support Vector Regression (SVR), and Decision Tree models perform best with near-perfect results after hyperparameter adjustment. The MLP Regressor and SVR models consistently exhibit high accuracy, low errors, and strong coefficients of determination, making them viable candidates for soil composition prediction. The Decision Tree model, while slightly less accurate than the top-performing models, has the advantage of being interpretable, providing a clear comprehension of the decision-making procedure. This interpretability can be useful in assessing the links between soil parameters and composition, assisting with domain knowledge extraction and decision-making processes. Ensemble models, such as Random Forest, also perform well, especially after hyperparameter tweaking. However, their increased complexity may sacrifice interpretability for greater forecast accuracy.

It is vital to highlight that the most relevant model should be chosen based on criteria such as accuracy requirements, computing resources, interpretability requirements, and soil composition data specifics. Furthermore, additional validation and testing on independent datasets is advised to confirm the universality and resilience of the chosen models. Finally, this work demonstrates the accuracy and reliability of soil composition prediction using machine learning approaches in conjunction with appropriate hyperparameter optimization methodologies. The findings benefit soil scientists, agricultural researchers, and environmental monitoring applications by allowing for more informed decision-making and a better knowledge of soil attributes.

References

Ana, C., Câmara, F., Marcos, B., Ceddia, Elias, M., Costa, É., Flávia, M., Pinheiro, M., Melo, D., Nascimento, G., and Vasques, M., 2022. Use of Airborne Radar Images and Machine Learning Algorithms to Map Soil Clay, Silt, and Sand Contents in Remote Areas under the Amazon Rainforest. Remote sensing, 14(22): pp. 5711-5711.

Bergstra, J. and Bengio, Y., 2012. Random Search for Hyper-Parameter Optimization. J. Mach. Learn. Res., 13, pp. 281-305. https://doi.org/10.5555/2503308.2188395.

Bousbih, S., Zribi, M., Pelletier, C., Gorrab, A., Lili-Chabaane, Z., Baghdadi, N.N., Aissa, N.B. and Mougenot, B., 2019. Soil Texture Estimation Using Radar and Optical Data from Sentinel-1 and Sentinel-2. Remote. Sens., 11, 1520.

Chagas, C., Junior, W., Bhering, S. and Filho, B., 2016. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena, 139, 232-240. https://doi.org/10.1016/J.CATENA.2016.01.001.

Ferreira, A.C.D.S., Ceddia, M.B., Costa, E.M., Pinheiro, É.F., Nascimento, M.M.D. and Vasques, G.M., 2022. Use of Airborne Radar Images and Machine Learning Algorithms to Map Soil Clay, Silt, and Sand Contents in Remote Areas under the Amazon Rainforest. Remote Sensing, 14(22), 5711.

Forkuor, G., Hounkpatin, O., Welp, G. and Thiel, M., 2017. High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. PLoS ONE, 12. https://doi.org/10.1371/journal.pone.0170478.

Gomez, C., Dharumarajan, S., Féret, J.-B., Lagacherie, P., Ruiz, L. and Sekhar, M., 2019. Use of Sentinel-2 Time-Series Images for Classification and Uncertainty Analysis of Inherent Biophysical Property: Case of Soil Texture Mapping. Remote Sensing, 11(5), 565. https://doi.org/10.3390/rs11050565.

Keshavarzi, A., Árbol, M., Kaya, F., Gyasi-Agyei, Y. and Rodrigo-Comino, J., 2022. Digital mapping of soil texture classes for efficient land management in the Piedmont plain of Iran. Soil Use and Management, 38, pp. 1705 - 1735. https://doi.org/10.1111/sum.12833.

Khalil, R.Z., Khalid, W. and Akram, M., 2016. Estimating of soil texture using landsat imagery: A Case Study of Thatta Tehsil, Sindh. 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 3110–3113. https://doi.org/10.1109/IGARSS.2016.7729804

Laborczi, A., Szatmári, G., Takács, K. and Pásztor, L., 2015. Mapping of Topsoil Texture in Hungary Using Classification Trees. Journal of Maps, pp. 1–11. https://doi.org/10.1080/17445647.2015.1113896

Liang, S., Meiyu, Z., and Bing, W., 2023. Predictive Soil Mapping Based on the Similarity of Environmental Covariates Using a Spatial Convolutional Autoencoder. Soil Science Society of America Journal, 87(3): pp. 631-643.

Maino, A., Alberi, M., Anceschi, E., Chiarelli, E., Cicala, L., Colonna, T. and Strati, V., 2022. Airborne Radiometric Surveys and Machine Learning Algorithms for Revealing Soil Texture. Remote Sensing, 14(15), 3814.

Mohammed, D., Ghada, A., Souham, M., and Walaa, A., 2023. Deep Learning-Based Framework for Soil Moisture Content Retrieval of Bare Soil from Satellite Data. Remote sensing, 15(7): pp. 1916-1916.

Naimi, S., Ayoubi, S., Demattê, J., Zeraatpisheh, M., Amorim, M. and Mello, F., 2021. Spatial Prediction of Soil Surface Properties in an Arid Region Using Synthetic Soil Image and Machine Learning. Geocarto International, 37, pp. 8230-8253. https://doi.org/10.1080/10106049.2021.1996639.

Niang, M., Nolin, M., Jégo, G. and Perron, I., 2014. Digital Mapping of Soil Texture Using RADARSAT-2 Polarimetric Synthetic Aperture Radar Data. Soil Science Society of America Journal, 78, pp. 673-684. https://doi.org/10.2136/SSSAJ2013.07.0307.

Rengma, N.S., Yadav, M., Kalambukattu, J.G. and Kumar, S., 2023. Machine Learning-Based Digital Mapping of Soil Organic Carbon and Texture in the MID-HIMALAYAN TERRAIN. Environmental Monitoring and Assessment, 195(8), 994. https://doi.org/10.1007/s10661-023-11608-9.

Vladyslav, Y., Andrii, P., Heorhii, K., 2022. Convolutional Neural Network Hyperparameter Optimization Applied to Land Cover Classification. Radìoelektronnì ì komp'ûternì sistemi, pp. 115-128.

Wadoux, A., Padarian, J. and Minasny, B., 2018. Multi-Source Data Integration for Soil Mapping Using Deep Learning. SOIL. https://doi.org/10.5194/SOIL-5-107-2019.

Yasser, A., Ali, E., Mahrous, A., Al-Razgan, M., and Ali, M., 2023. Hyperparameter Search for Machine Learning Algorithms for Optimizing the Computational Complexity. Processes, 11(2): p. 349-349.