KIMYO INTERNATIONAL UNIVERSITY IN TASHKENT
CENTRAL ASIAN JOURNAL OF STEM
ISSN 2181–2934 https://stem.kiut.uz
Vol. 5, Issue 4, Dec, 2025
https://doi.org/10.5281/zenodo.18654321

AI-Driven Soil Organic Carbon Prediction Using Random Forest: A Data-Driven Study on Uzbekistan's Agricultural Soils

1st Abdurasul Bobonazarov, *Department of Control and Computer Engineering, Turin Polytechnic University in Tashkent*
Tashkent, Uzbekistan, a.bobonazarov@polito.uz
2nd Akmal Rustamov, *Department of Control and Computer Engineering, Turin Polytechnic University in Tashkent*
Tashkent, Uzbekistan,
3rd Khamidulla Khabibullaev, *Department of Control and Computer Engineering, Turin Polytechnic University in Tashkent*
Tashkent, Uzbekistan,

Abstract

Soil Organic Carbon (SOC) is a key determinant of soil health and agricultural sustainability. However, large-scale assessment remains challenging because of the costly and labor-intensive laboratory procedures. This challenge is particularly relevant to Uzbekistan, where soil degradation and nutrient depletion threaten long-term productivity. To address the need for scalable SOC estimation, this study evaluated whether machine learning, specifically a Random Forest (RF) model, can accurately predict SOC using only low-cost agrochemical data and geospatial information. Unlike many existing studies that depend either on extensive physicochemical soil profiles or spatially rich environmental covariates, this study introduces a hybrid minimal-feature approach that combines basic laboratory measurements with approximated sampling coordinates. This study aimed to develop, optimize, and evaluate an RF-based SOC prediction model using a comprehensive dataset of 97,449 soil samples collected between 2022 and 2024. The document outlines the methodological workflow, including data preprocessing, correlation analysis, baseline modeling, and hyperparameter optimization using GridSearchCV.

The optimized RF model achieved an R^2 of 0.619 and an RMSE of 0.243 on the test set, outperforming the baseline configuration and demonstrating a stable predictive behavior across most SOC values. These results show that meaningful SOC estimation is possible even in data-limited contexts, marking the first large-scale AI-driven SOC prediction study based on nationally collected soil laboratory data from Uzbekistan. These findings highlight a practical pathway for developing digital soil monitoring tools in regions with sparse

environmental datasets. Future research should incorporate temporal indicators, additional soil attributes, and remote sensing features to further enhance model accuracy and support advanced spatiotemporal soil analytics.

Index Terms

Humus Prediction, Machine Learning, Random Forest, Support Vector Regressor, Agrochemical Analysis, Precision Agriculture, Fertilizer Optimization, AI-Driven Soil Analysis, Data Analytics, Smart Farming.

I. Introduction

Soil Organic Carbon (SOC) is a critical indicator of soil health, influencing nutrient availability, microbial activity, and long-term agricultural productivity. Monitoring SOC is fundamental for mitigating land degradation and promoting sustainable farming practices, particularly in arid regions such as Uzbekistan, where declining soil fertility poses significant challenges to crop production. Traditional SOC measurements rely on laboratory analysis, which is accurate but costly, labor-intensive, and unsuitable for high-resolution or large-scale monitoring. These limitations have accelerated interest in data-driven approaches capable of providing rapid and scalable SOC estimations.

Advances in machine learning have enabled substantial progress in digital soil mapping by modeling nonlinear relationships among soil, environmental variables, and geospatial variables. Ensemble-based algorithms, such as Random Forest and Gradient Boosting, have consistently demonstrated strong predictive capabilities compared to classical linear methods [1]–[3]. Previous global and regional studies have reported SOC prediction performance with R^2 values typically ranging between 0.55 and 0.75, depending on feature richness and spatial heterogeneity [4], [5]. The increasing availability of soil datasets, together with improvements in computational methods, has encouraged the integration of machine learning into precision agriculture workflows.

This study also builds upon our earlier contribution to AI-driven agricultural analytics, particularly the NAS-GBM model for crop yield prediction introduced in [6]. That study demonstrated the effectiveness of automated model design in agronomic prediction tasks, motivating the extension of AI methodologies from crop-level forecasting to soil quality assessment. In this context, the present study evaluated whether a Random Forest-based approach can provide accurate SOC predictions using a minimal set of nutrient and geospatial inputs.

While numerous studies have applied machine learning for SOC prediction, most rely either on spatially explicit environmental covariates, often derived from remote sensing, or on extensive soil laboratory measurements that include nitrogen content, pH, salinity, texture, and other physico-chemical attributes [1]–[5]. In contrast, this study introduces a hybrid minimal-feature approach that integrates basic agrochemical indicators (P_2O_5, K_2O, and humus) with geospatial coordinates to produce SOC predictions using only low-cost data. The fusion of spatial and chemical information offers a practical and scalable modeling strategy for data-limited regions. To the best of our knowledge, this study is the first large-scale AI-driven investigation of SOC prediction using nationally collected soil laboratory data from Uzbekistan, thereby addressing a significant regional research gap and contributing new insights to digital soil mapping efforts in Central Asia.

II. Materials and Methods

The dataset used in this study consists of soil samples submitted by farmers for routine agrochemical analysis. Each sample included measurements of P_2O_5, K_2O, and humus (%), the latter of

which was used in this study as a proxy for Soil Organic Carbon (SOC). This substitution follows the standard agronomic practice in Uzbekistan, where humus content is commonly reported and operationally treated as an indicator of soil organic matter and SOC levels.

Additional soil physicochemical parameters, such as nitrogen content, soil pH, electrical conductivity (salinity), moisture levels, texture composition, cation exchange capacity, and micronutrient concentrations, were not included in the dataset, likely due to higher laboratory analysis costs or limited testing requests from farmers. Furthermore, the dataset does not contain timestamps for sample collection or laboratory processing, preventing the incorporation of seasonal and temporal variability into the analysis. In cases where precise geographic coordinates of sampling were unavailable, the centroid of the respective district was used to approximate the sample location.

Despite these constraints, the dataset comprises 97,449 soil samples collected nationwide between 2022 and 2024, providing a large-scale and regionally representative basis for machine-learning-based SOC prediction. The combination of humus measurements with geospatial coordinates offers a practical approach for modeling SOC in data-limited environments, such as Uzbekistan.

A baseline Random Forest model was trained using the default parameters. Hyperparameter tuning was performed using GridSearchCV, exploring variations in the number of estimators, maximum depth, minimum split size, minimum leaf size, and feature selection strategies.

III. Results

A correlation assessment was conducted to examine the relationships among the available variables before model development. The analysis showed that SOC is moderately associated with geospatial attributes, with a correlation of approximately 0.42 and 0.47 with latitude and longitude, respectively (Fig. 1). These values indicate that spatial gradients play a substantial role in shaping SOC variability in Uzbekistan. In contrast, the nutrient indicators P_2O_5 and K_2O displayed much weaker correlations with SOC, suggesting that their influence is either nonlinear or mediated by additional soil properties not captured in this dataset. These observations justify the use of a nonlinear modeling technique, such as Random Forest, to better represent the underlying relationships.

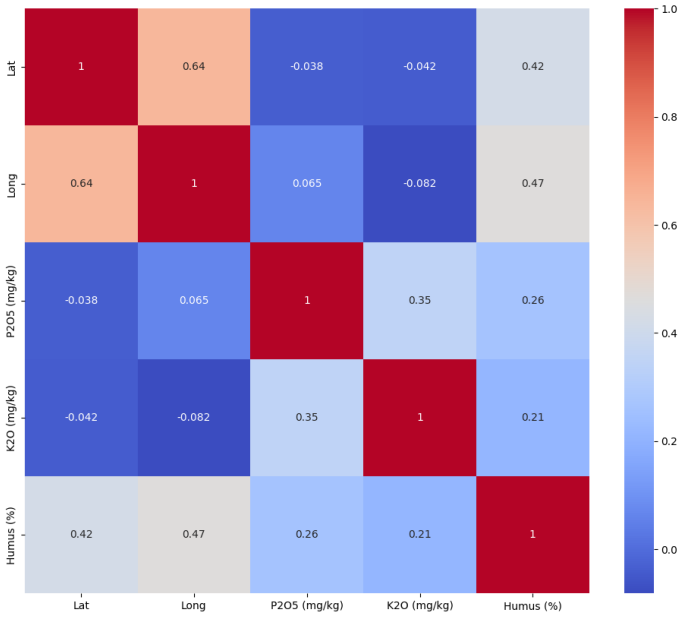

Fig. 1. Correlation heatmap of SOC and predictor variables (P_2O_5, K_2O, latitude, longitude).

A comparative evaluation of the baseline and tuned Random Forest models is presented in Table 1. The baseline model achieved an R^2 of 0.577 and an RMSE of 0.256 on the test set, indicating moderate predictive accuracy. After hyperparameter optimization, the tuned model achieved an improved R^2 of 0.619 and a reduced RMSE of 0.243. The corresponding training metrics show that tuning effectively decreased the gap between the training and testing performances, thereby mitigating overfitting.

Although Table 1 summarizes only the performance metrics, the optimized model was obtained using a configuration that included a larger number of trees, a moderately deep tree structure, and stricter constraints on node splitting. Specifically, the best-performing Random Forest model used 400 estimators, a maximum depth of 30, a minimum of 5 samples required to split an internal node, a minimum of 2 samples per leaf, and 'sqrt' as the feature selection strategy at each split. These parameter choices increased the model robustness while improving generalization compared to the baseline configuration.

TABLE I
BASELINE VS TUNED RANDOM FOREST PERFORMANCE

Model	**Dataset**	μ_e	σ_e	**MSE**	R^2
Baseline Random Forest	Training	0.5×10^{-3}	0.132	0.017	0.891
	Test	0.5×10^{-3}	0.257	0.066	0.577
Tuned Random Forest	Training	0.2×10^{-3}	0.196	0.039	0.760
	Test	0.7×10^{-3}	0.244	0.059	0.619

The predictive behavior of the optimized Random Forest model is illustrated in Fig.2, which compares the predicted SOC values with their corresponding observed measurements. Most points fell close to the 1:1 reference line, indicating a strong agreement within the dominant SOC range of 0-2%. This suggests that the model was well calibrated for the typical soils found in the dataset. A slight deviation from the reference line was noticeable for samples with higher SOC values, reflecting a mild tendency toward underprediction in this upper range. This behavior is frequently observed in ensemble models applied to datasets with skewed target distributions, where high-end values are comparatively scarce. Overall, the scatter plot demonstrates that the tuned model captures the main structure of SOC variability with a reasonable fidelity.

Fig. 3 presents the distribution of the residual errors for the tuned Random Forest model. The residuals exhibited an approximately symmetric, bell-shaped pattern centered around zero, indicating that the model did not systematically overestimate or underestimate SOC across the dataset. The relatively narrow error spread suggests consistent predictive performance, with most deviations falling within a small range. This distribution supports the numerical evaluation and confirms that the model achieves a balanced bias – variance trade-off. The absence of noticeable skewness or extreme outliers further reinforces the model stability and suitability for SOC prediction in this dataset.

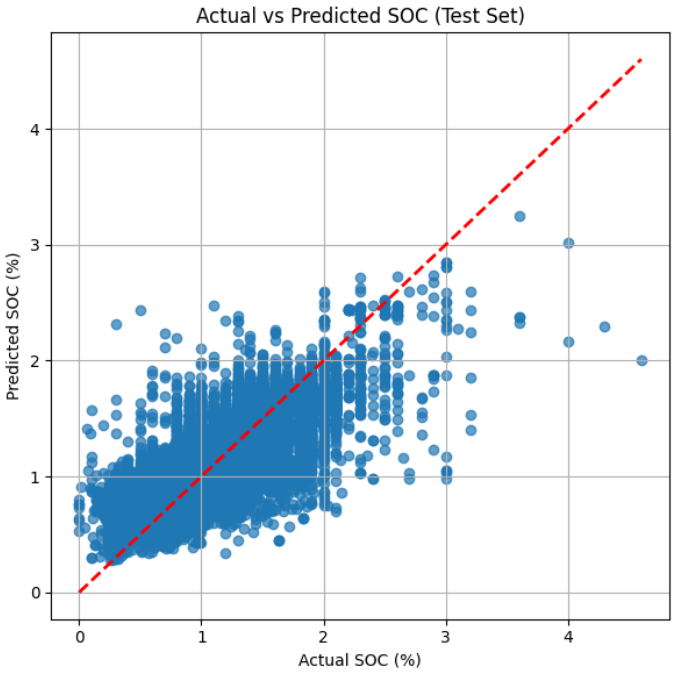

Fig. 2. Predicted vs. actual SOC value for the test dataset.

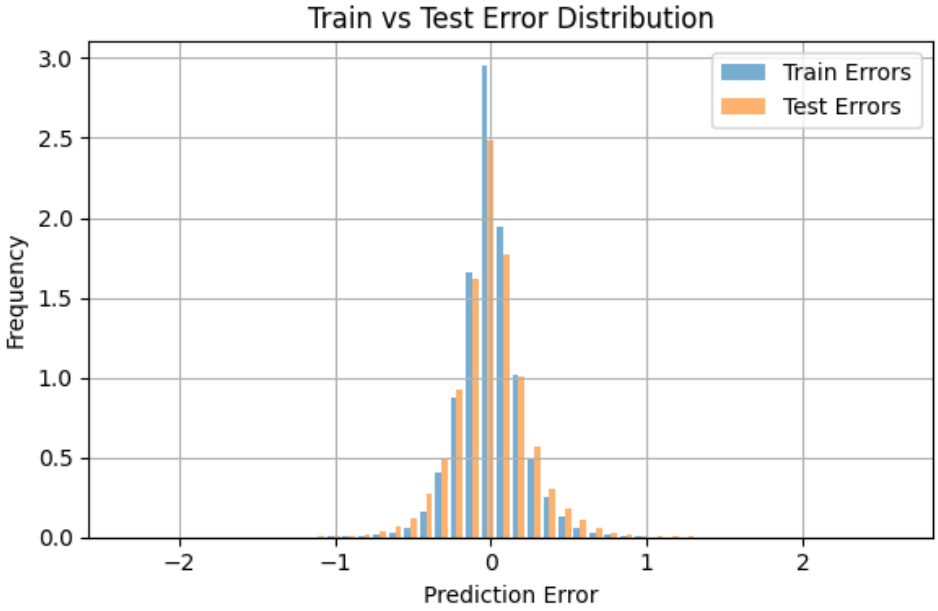

Fig. 3. Residual error distribution for the tuned Random Forest model.

The dataset lacks key soil attributes, such as nitrogen, pH, salinity, moisture, and texture, limiting the explanatory power of the model. Missing timestamps prevent the modeling of seasonal variability. The use of district centroids may introduce spatial approximation errors.

IV. Conclusions

This study demonstrates that the Random Forest algorithm can serve as an effective and practical approach for predicting Soil Organic Carbon (SOC) in Uzbekistan using a limited set of low-cost in-

put features. Despite relying only on basic agrochemical measurements and geospatial coordinates, the tuned model achieved performance levels comparable to those reported in broader digital soil mapping studies. These results highlight the potential of deploying machine learning tools to support soil monitoring and decision-making in regions where comprehensive environmental datasets are not readily available. This study also provides a much-needed baseline for AI-driven soil analysis in Uzbekistan, establishing a foundation for future digital agriculture initiatives in the country.

Several avenues exist for enhancing SOC prediction accuracy. Incorporating a wider range of soil physicochemical parameters, such as nitrogen, pH, salinity, moisture content, and texture, would enrich the feature space and allow the models to capture more of the underlying variability. The inclusion of temporal information, such as sampling dates or seasonal indicators, could further enable the development of models that reflect the dynamic soil processes. Integrating remote sensing data, IoT-based soil sensors, and advanced spatial–temporal learning architectures may also substantially improve prediction performance. Collectively, these extensions would facilitate the creation of more accurate and scalable digital soil assessment tools to support sustainable land management and agricultural development in Uzbekistan.

V. Acknowledgements

The authors express their gratitude to the Republican Center for Agrochemical Analysis of Uzbekistan for providing access to soil laboratory data used in this study.

References

[1] Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez Gonzalez M, Kilibarda M, et al., SoilGrids250m: Global gridded soil information based on machine learning. PLoS One, vol. 12, no. 2, pp. 1-40, (2017). https://doi.org/10.1371/journal.pone.0169748

[2] A.M.J.C. Wadoux, B. Minasny, and A.B. McBratney. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, vol. 210, 2020. https://doi.org/10.1016/j.earscirev.2020.103359

[3] S. van der Westhuizen, G.B.M. Heuvelink and D.P. Hofmeyr, Multivariate random forest for digital soil mapping Author links open overlay panel. Geoderma, vol. 431, 2023. https://doi.org/10.1016/j.geoderma.2023.116365

[4] A.B. McBratney, M.L. Mendonça Santos, and B. Minasny. On digital soil mapping. Geoderma, vol. 117 no. 1-2, 2003. https://doi.org/10.1016/S0016-7061(03)00223-4

[5] R. Dorijan, D. Jug, I. Jug, and M. Jurišić. A Comprehensive Evaluation of Machine Learning Algorithms for Digital Soil Organic Carbon Mapping on a National Scale. Applied Sciences, vol. 14, no. 21, (2024). https://doi.org/10.3390/app14219990

[6] S. Anakal, A. Bobonazarov, J.V. Naga Ramesh, E. Muniyandy, M. Manjusha, Y.A. Baker El-Ebiar. AI-Driven NAS-GBM Model for Precision Agriculture: Enhancing Crop Yield Prediction Accuracy. International Journal of Advanced Computer Science and Applications, vol. 16, no. 3, (2024). https://doi.org/10.14569/IJACSA.2025.0160373

[7] Abdivakhidov, K. (2023). Application and removing protective metal coatings. *AIP Conference Proceedings*, 2789, 040087. https://doi.org/10.1063/5.0149617

[8] Abdivakhidov, K., & Sharipov, K. (2024). Innovative metal coating Technologies for enhanced corrosion protection: A Comprehensive review of advanced solutions. *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.15696104

[9] Abdivakhidov, K., & Sharipov, K. (2024). Corrosion-resistant protective coatings for metals: A review of metallic and non-metallic coatings. *AIP Conference Proceedings, 3045*, 060011. https://doi.org/10.1063/5.0197373