Predicting Subsurface Soil Parameters Using Surface and Satellite Data with Machine Learning Techniques

Andrea Luces, Jean Rendon, Afshin Gholamy, and Leobardo Valera

Abstract Subsurface soil properties such as porosity, permeability and saturation are critical for applications in agriculture, water resource management, and natural resource exploration. However, direct measurement of these parameters is often costly, labor-intensive, and spatially constrained, making it impractical for large-scale assessments. To address these limitations, we present a machine learning-based framework for predicting subsurface soil characteristics by leveraging a fusion of surface-level observations and satellite-derived data. Our approach integrates processing heterogeneous data sources, including remote sensing products (e.g., Sentinel-1, Sentinel-2, and SMAP), surface soil measurements, and topographic features. These inputs are processed through a feature extraction and transformation pipeline, which feeds into multiple predictive models. We evaluate several machine learning techniques including Kernel K-Nearest Neighbors (KKNN), Random Forests (RF), Neural Networks, and Least Squares Regression to identify the most effective strategies for accurate soil parameter estimation. Furthermore, these models can be utilized to characterize the various subsurface layers and to visualize potential mineral reserves, particularly hydrocarbons. By integrating well data (logs, production, etc), seismic information available, and other geological informa-

Andrea Luces
El Paso Community College, 919 Hunter Dr
El Paso, TX 79915, USA
e-mail: andrealuces@gmail.com

Jean Rendon
El Paso Community College, 919 Hunter Dr, El Paso, TX 79915, USA
e-mail: jeanrendon@gmail.com

Afshin Gholamy
National University, 9388 Lightwave Ave, San Diego, CA 92123, USA
e-mail: agholamy@nu.edu

Leobardo Valera
El Paso Community College, 919 Hunter Dr, El Paso, TX 79915, USA, USA
e-mail: lvalerav@epcc.edu

tion into the models, it becomes possible to identify zones with a high potential for hydrocarbon accumulation, especially in under-explored or poorly studied areas.

1 Background and Related Work

Reliable soil data at varying depths plays a critical role in a wide range of applications, including sustainable land management, precision agriculture, groundwater modeling, and geotechnical engineering. Despite their importance, acquiring accurate subsurface soil measurements remains a significant challenge due to the high costs, limited spatial coverage, and logistical demands associated with traditional field sampling methods.

In response to these challenges, recent research has turned to remote sensing and data-driven approaches for estimating subsurface properties. One such effort is the Soil Moisture Sensing Performance Indicators and Evaluation (**SOMOSPIE**) project, which demonstrated the feasibility of enhancing surface soil moisture estimation by combining in-situ measurements with *machine learning techniques* and satellite-based remote sensing data. [1, 2, 3].

This study aims not only to analyze surface characteristics but also to characterize the various subsurface layers through the application of machine learning algorithms. The ability to estimate key petrophysical properties—such as porosity, permeability, and fluid saturations (oil, water, and/or gas)—using existing well logs presents a cost-effective and efficient alternative for evaluating new exploration prospects, especially in data-scarce environments.

By extracting fundamental reservoir properties from available well logs, the project seeks to build a robust petrophysical model. These results will then be integrated with any additional geological information to strengthen the regional geological characterization and increase the likelihood of identifying new hydrocarbon-bearing prospects.

Machine learning (ML) offers a powerful solution for this problem taking surface-level proxies as input to infer subsurface conditions. Studies have shown that algorithms like Random Forests (RF), Neural Networks, and K-Nearest Neighbors (KNN) can effectively model complex nonlinear relationships in spatial and environmental data [4, 5]. Additionally, least squares models provide interpretable baseline predictors for geophysical data [6]. Recently, the AgroLens project was used to predict soil nutrients such as phosphorus, potassium, and nitrogen using the LUCAS soil dataset and Sentinel-2 imagery, without relying on lab-based testing [7]. Their methodology supports the idea that satellite data can be reliably used to infer key soil parameters.

Other researchers used Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Multilayer Perceptrons to generate spatial maps of soil texture components (clay, silt, sand) and organic carbon content and evaluated their use in hydrological models [8]. These models use geomorphometric

features, and subsurface genetic horizons as input, showing that subsoil features can improve the spatial prediction of soil classes [9].

A recent study used a hybrid model combining Convolutional Neural Networks (CNNs) Long Short-Term Memory (LSTM) (ConvLSTM) model to combine Soil Moisture Active Passive (SMAP) satellite data, North American Land Data Assimilation System, version 2 (NLDAS-2) weather data, and Soil Landscapes of the United States at 100-meter resolution (SOLUS100) soil maps to produce daily estimates of surface and subsurface soil moisture. The integration of soil property maps (such as texture and bulk density) was critical to improving accuracy [10].

We propose a machine learning pipeline that integrates surface-level remote sensing inputs and geospatial features to predict subsurface soil properties using machine learning techniques (**Fig. 1**). This approach aims to reduce reliance on costly fieldwork while improving the spatial coverage of key subsurface indicators essential for agricultural and environmental decision-making. Our framework is designed to predict a broader range of parameters including porosity, permeability, and soil texture providing a more comprehensive understanding of subsurface conditions.

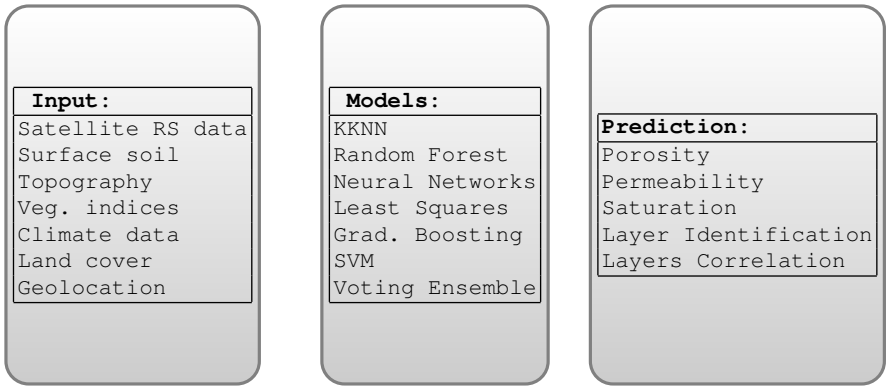

Fig. 1 Learning pipeline that integrates surface-level remote sensing inputs and geospatial features to predict subsurface soil properties using machine learning techniques

2 Metodology

The process of predicting subsurface soil parameters consists of four steps: data acquisition, data preprocessing, model training and prediction. In this preliminary research, the input and target data are collected from the U.S. Geological Survey's GeoLog Archiver, and we focus in the data of the wells located in the state of Texas (**Fig. 2**).

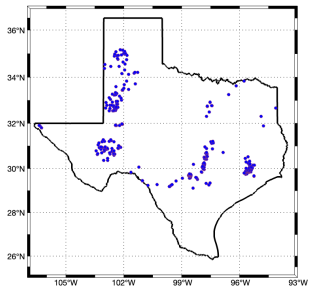

Fig. 2 Location of the wells where the preliminary data was extracted

To capture the relationship between the surface soil parameters and the subsurface parameters (e.g., porosity, permeability, moisture at depth), we train different supervised machine learning models such as: Kernel K-Nearest Neighbors (KKNN), Random Forest (RF), Neural Networks, and Least Squares Regression. Once each model is trained, we use them to predict the subsurface parameters. Subsequently, available geological information—such as sedimentological models and seismic studies—can be integrated to train the model, enabling the prediction of key reservoir parameters with greater accuracy in any desired location.

3 Conclusion and Future Work

We are developing a modular machine learning pipeline for predicting subsurface soil parameters using well log, any geological information available, surface-level observations and satellite data. As first step, we developed a modular pipeline to train KKNN [11, 12] model using existing data and we use such model to make some predictions.

The initial results have been modest, but the results are promising, so we are ready to start to implement the rest of the models.

As future work, we plan to expand this prototype into a fully integrated system named **DeepSoilMap: A Machine Learning Framework for Predicting Subsurface Parameters from**. This system will implement all evaluated models and include capabilities for real-time data ingestion, model selection, uncertainty quantification, and interactive visualization.

Acknowledgments

The authors gratefully acknowledge Desert Image (Diagnostic Imaging and Radiology in El Paso, TX), whose flexibility and encouragement were vital to the advancement of this work. We also thank the TRACS (Theoretical Research and its Application in Computer Science in The University of Texas at El Paso) group for granting us access to their research facilities, which played a crucial role in enabling the technical development of this study.

We further express our appreciation to the School of Arts, Letters and Sciences at National University for their continued encouragement. Special thanks are due to the Department of Mathematics and Natural Sciences, whose mentorship, resources, and collaborative atmosphere provided a strong foundation for the successful completion of this project.

References

1. D. Rorabaugh, M. Guevara, R. M. Llamas, J. Kitson, R. Vargas, and M. Taufer, "SOMOSPIE: A modular soil moisture spatial inference engine based on data-driven decisions", *Proceedings of the 2019 15th International Conference on eScience (eScience)*, IEEE, pp. 1–10, 2019.
2. R. M. Llamas, L. Valera, P. Olaya, M. Taufer, and R. Vargas, "Downscaling satellite soil moisture using a modular spatial inference framework", *Remote Sensing*, vol. 14, no. 13, Art. no. 3137, 2022.
3. R. Bindlish, T. J. Jackson, D. M. Le Vine, and M. H. Cosh, "Validation of SMAP soil moisture using core validation sites and the SOMOSPIE project", *Remote Sensing of Environment*, vol. 259, 2021.
4. G. Biau and E. Scornet, "A random forest guided tour", *TEST*, vol. 25, no. 2, pp. 197–227, 2016.
5. D. J. Lary, A. H. Alavi, A. Gandomi, and A. L. Walker, "Machine learning in geosciences and remote sensing", *Geoscience Frontiers*, vol. 7, no. 1, pp. 3–10, 2016.
6. Y. Yuan, X. Chen, and T. Wu, "A least-squares based approach for regional soil moisture estimation", *Journal of Hydrology*, vol. 576, pp. 356–367, 2019.
7. C. Kammerlander, V. Kolb, M. Luegmair, L. Scheermann, M. Schmailzl, M. Seufert, J. Zhang, D. Dalic, and T. Schön, "Machine Learning Models for Soil Parameter Prediction Based on Satellite, Weather, Clay and Yield Data", *arXiv preprint*, arXiv:2503.22276, 2025.
8. F. Alonso-Sarria, A. Blanco-Bernardeau, F. Gomariz-Castillo, H. Jimenez-Bastida, and A. Romero-Diaz, "Estimation of soil properties using machine learning techniques to improve hydrological modeling in a semiarid environment: Campo de Cartagena (Spain)", *Earth Science Informatics*, vol. 18, no. 3, 2025, pp. 1–24.
9. S. Manteghi, K. Moravej, S. R. Mousavi, M. A. Delavar, and A. Mastinu, "Digital soil mapping for soil types using machine learning approaches at the landscape scale in the arid regions of Iran", *Advances in Space Research*, vol. 74, no. 1, 2024, pp. 1–16.
10. S. Rabiei, E. Babaeian, and S. Grunwald, "Surface and Subsurface Soil Moisture Estimation Using Fusion of SMAP, NLDAS-2, and SOLUS100 Data with Deep Learning", *Remote Sensing*, vol. 17, no. 4, 2025, Art. no. 659.
11. A. Gholamy, J. Parra, V. Kreinovich, O. Fuentes, and E. Anthony, "How to best apply deep neural networks in geosciences: Towards optimal 'averaging' in dropout training", In: *Junzo Watada, Shing Chieng Tan, Pandian Vasant, Eswaran Padmanabhan, and Lakhmi C. Jain (eds.), Smart Unconventional Modelling, Simulation and Optimization for Geosciences and Petroleum Engineering*, Springer Verlag, 2019, pp. 15-26.

12. A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation", *International Journal of Intelligent Technologies & Applied Statistics*, 2018, Vol. 11, No. 2, pp. 105-111.