Soil Moisture Prediction Using Machine Learning: An End-to-End Regression Approach Using Sentinel-1 SAR and SMAP Data

Sachin Dimri

Indian Institute of Technology Bombay

A Machine Learning and Deep Learning Approach

Case Study Submission – Vassar Labs IT Solutions

Job Code 2 – Data Science

February 12, 2026

Abstract

Soil moisture is a critical parameter in agriculture, hydrology, and climate monitoring. This study develops an end-to-end machine learning pipeline to predict soil moisture content using Sentinel-1 SAR backscatter coefficients (VV, VH) and NASA SMAP satellite data. The dataset comprises 30,747 observations with three predictor variables and one continuous target variable. Multiple regression algorithms were systematically evaluated, including Linear Regression, Polynomial Regression, Ridge Regression, Random Forest, Gradient Boosting, Support Vector Regression (SVR), and Deep Neural Networks. Data preprocessing involved outlier detection and removal, exploratory data analysis, feature scaling, and rigorous assumption checking. Results indicate that satellite-derived radar backscatter features exhibit minimal predictive signal for soil moisture ($R^2 \approx 0.01$–0.06), with SVR using RBF kernel and Neural Networks achieving the highest performance ($R^2 = 0.053$–0.056). The study concludes that additional environmental predictors such as temperature, rainfall, soil type, and temporal features are necessary to substantially improve prediction accuracy. This work provides a complete, reproducible machine learning workflow suitable for academic evaluation and demonstrates proper scientific methodology in addressing data limitations.

Keywords: Soil moisture prediction, Sentinel-1 SAR, SMAP satellite, machine learning, regression modeling, remote sensing, deep learning

Contents

1 Introduction

1.1 Background

Soil moisture represents the water content present in the top layers of soil and plays a fundamental role in the Earth's hydrological cycle. It directly influences agricultural productivity, water resource management, climate variability, and ecosystem dynamics [1,2]. Traditional ground-based soil moisture measurement methods using in-situ sensors are expensive, time-consuming, and geographically limited, making large-scale monitoring challenging.

Satellite-based remote sensing offers a cost-effective solution for continuous, large-scale soil moisture monitoring. Two prominent satellite missions provide valuable data for this purpose: **Sentinel-1**, operated by the European Space Agency (ESA), delivers Synthetic Aperture Radar (SAR) backscatter coefficients that are sensitive to surface moisture and roughness; and **SMAP (Soil Moisture Active Passive)**, a NASA mission, provides direct soil moisture measurements using L-band radiometry [3,4].

1.2 Motivation

The motivation behind this project stems from the increasing global need for:

1. **Precision Agriculture** – Optimizing irrigation schedules and water usage efficiency to improve crop yields while conserving water resources.
2. **Efficient Irrigation Management** – Supporting farmers in making data-driven decisions about when and where to irrigate.
3. **Drought Monitoring** – Early identification of low-moisture regions to enable preventive measures and disaster preparedness.
4. **Climate Change Adaptation** – Understanding soil-atmosphere interactions and improving climate models.
5. **Water Resource Management** – Assisting governmental agencies in sustainable groundwater and surface water management.

Accurate soil moisture estimation directly helps:

- Farmers optimize water usage and reduce irrigation costs.
- Governments manage water resources effectively.
- Scientists study climate variability and land-atmosphere interactions.
- Agricultural planners improve crop yield predictions and food security assessments.

From a data science perspective, this project provides hands-on experience in:

- Regression modeling techniques.
- Feature importance analysis.
- Model comparison and selection.
- Hyperparameter tuning.
- Production-ready ML pipeline design.
- Handling real-world data limitations.

1.3 Problem Statement

While satellite-based remote sensing provides large-scale coverage, translating raw radar backscatter signals into accurate soil moisture estimates requires advanced modeling techniques. The fundamental challenge addressed in this study is:

> **How can machine learning techniques be used to accurately predict soil moisture content from satellite radar backscatter data (VV, VH polarizations) combined with SMAP observations?**

This project addresses this challenge by designing a complete machine learning workflow encompassing data preprocessing, exploratory analysis, model development, evaluation, and interpretation.

1.4 Objectives

The primary objective of this project is to develop an end-to-end machine learning model to accurately predict soil moisture content using Sentinel-1 SAR backscatter coefficients and SMAP satellite data. The goal is to build a robust regression pipeline that analyzes remote sensing data and estimates soil moisture efficiently and reliably.

Specific objectives include:

1. Perform comprehensive exploratory data analysis (EDA) to understand data distributions, correlations, and potential issues.
2. Compare multiple machine learning algorithms for regression modeling.
3. Evaluate model performance using appropriate regression metrics (RMSE, MAE, R^2).
4. Identify the most influential features affecting soil moisture prediction.

5. Check and validate linear regression assumptions systematically.

6. Build a deployment-ready prediction pipeline.

7. Provide scientific interpretation of model limitations and data characteristics.

1.5 Use Cases

This soil moisture prediction system has several practical applications:

Precision Agriculture Predict soil moisture levels to guide irrigation scheduling, optimize water application rates, and improve crop productivity through targeted water management.

Drought Monitoring Identify low-moisture regions early to enable preventive measures, support drought early warning systems, and inform water allocation policies during water scarcity.

Flood Risk Assessment High soil moisture content can indicate saturation conditions that increase flood-prone risk, supporting disaster preparedness and emergency response planning.

Water Resource Management Assist authorities in managing groundwater recharge, reservoir operations, and irrigation system efficiency at regional and national scales.

Climate Research Enable researchers to study land-atmosphere interactions, validate hydrological models, and improve understanding of climate change impacts on water cycles.

2 Literature Review

2.1 Remote Sensing for Soil Moisture

Synthetic Aperture Radar (SAR) remote sensing has emerged as a powerful tool for soil moisture estimation due to its sensitivity to dielectric properties of soil and its all-weather, day-and-night observation capabilities [5]. The Sentinel-1 mission, part of the European Union's Copernicus Programme, provides free and open C-band SAR data with high temporal resolution (6–12 day revisit time) and fine spatial resolution (10–20 meters) [6].

Research has demonstrated that Sentinel-1 backscatter coefficients, particularly VV (Vertical-Vertical) and VH (Vertical-Horizontal) polarizations, correlate with surface soil moisture conditions. VV polarization is more sensitive to surface scattering and soil

moisture, while VH polarization captures volume scattering and vegetation structure effects [7,8].

NASA's SMAP mission, launched in 2015, measures soil moisture in the top 5 cm of soil globally every 2–3 days using L-band radiometry (1.4 GHz frequency) [9]. SMAP data provides coarser spatial resolution (approximately 36 km) but offers direct soil moisture measurements that complement the high-resolution SAR data.

2.2 Machine Learning Applications

Machine learning techniques have shown promise in soil moisture prediction by capturing complex non-linear relationships between radar signals and soil water content. Recent studies have explored various algorithms:

- **Random Forest and ensemble methods** have been successfully applied to integrate multiple remote sensing features and environmental variables [10].
- **Support Vector Machines (SVM)** with non-linear kernels can capture complex decision boundaries in soil moisture retrieval [11].
- **Deep learning approaches**, particularly Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs), have demonstrated improved accuracy when sufficient training data is available [12].
- **Hybrid approaches** combining physical models with machine learning have shown potential for improving generalization across different landscapes and seasons [13].

2.3 Research Gaps

Despite progress in satellite-based soil moisture retrieval, several challenges remain:

1. **Vegetation effects** – Dense vegetation attenuates radar signals, complicating soil moisture retrieval in agricultural and forested areas.
2. **Surface roughness** – Soil surface roughness influences backscatter independently of moisture content.
3. **Limited predictors** – Using only radar backscatter without auxiliary environmental data (temperature, rainfall, soil type) limits prediction accuracy.
4. **Temporal dynamics** – Most studies focus on spatial patterns rather than temporal evolution of soil moisture.
5. **Model interpretability** – Black-box machine learning models often lack physical interpretability needed for scientific understanding.

This study contributes by systematically evaluating multiple machine learning approaches, rigorously checking model assumptions, and providing transparent analysis of model limitations when predictive signal is weak.

3 Dataset Description

3.1 Data Source

The dataset used in this project is derived from satellite-based remote sensing observations, specifically combining data from two satellite missions.

Sentinel-1 SAR (Synthetic Aperture Radar)

- **Operator:** European Space Agency (ESA) as part of the Copernicus Programme.
- **Sensor Type:** C-band Synthetic Aperture Radar operating at 5.405 GHz frequency.
- **Polarizations:** VV (Vertical transmit – Vertical receive) and VH (Vertical transmit – Horizontal receive).
- **Spatial Resolution:** 10–20 meters.
- **Temporal Resolution:** 6–12 days revisit time.
- **Observation Time:** Morning overpass (approximately 6:00 AM local time).
- **Advantages:** All-weather capability, day-and-night operation, sensitive to soil moisture and surface roughness.

Sentinel-1 provides radar backscatter coefficients expressed in decibels (dB). Radar signals interact with the land surface, and the reflected energy (backscatter) depends on dielectric properties (influenced by water content), surface roughness, and vegetation cover. VV polarization is particularly sensitive to surface scattering from soil, while VH polarization captures cross-polarized scattering influenced by vegetation structure.

SMAP (Soil Moisture Active Passive)

- **Operator:** NASA (National Aeronautics and Space Administration).
- **Mission Objective:** Global soil moisture and freeze/thaw state monitoring.
- **Sensor Type:** L-band radiometer (1.4 GHz frequency).
- **Measurement Depth:** Top 5 cm of soil.

- **Spatial Resolution:** Approximately 36 km.
- **Temporal Resolution:** 2–3 days global revisit.
- **Launch Date:** January 31, 2015.

SMAP measures naturally emitted microwave radiation from Earth's surface. The L-band frequency is ideal because atmospheric effects are minimal, and the signal penetrates through moderate vegetation to sense soil conditions. SMAP provides direct soil moisture estimates (volumetric water content, m^3/m^3).

Data Integration The dataset combines high spatial resolution Sentinel-1 radar backscatter with SMAP soil moisture observations. This integration leverages the complementary strengths of both missions: Sentinel-1's fine spatial detail and SMAP's direct moisture sensitivity. The data represents morning observations to capture relatively stable diurnal soil moisture conditions.

3.2 Dataset Characteristics

The dataset is structured in tabular format suitable for supervised regression machine learning.

Dataset Statistics:

- **Total Observations:** 30,747 records (after outlier removal: 30,742).
- **Number of Features:** 3 independent variables (predictors).
- **Target Variable:** 1 continuous dependent variable (`soil_moisture`).
- **Missing Values:** 0 (complete dataset with no missing data).
- **Duplicate Rows:** 0.
- **Memory Size:** 1.2 MB.
- **Data Types:** All numerical (float64).

Variable	Description
VV	Sentinel-1 radar backscatter coefficient in VV polarization (Vertical transmit – Vertical receive), measured in decibels (dB). Sensitive to surface scattering from soil moisture and surface roughness.
VH	Sentinel-1 radar backscatter coefficient in VH polarization (Vertical transmit – Horizontal receive), measured in decibels (dB). Sensitive to volume scattering from vegetation and cross-polarized surface interactions.
smap_am	SMAP morning soil moisture measurement (volumetric water content), dimensionless ratio between 0 and 1 representing the fraction of soil volume occupied by water.
soil_moisture	Target variable representing actual soil moisture content. This is the dependent variable to be predicted using machine learning models.

Table 1: Feature descriptions and characteristics

Feature Descriptions:

3.3 Variable Statistics

Variable	Mean	Std Dev	Min	Max
VV	-9.196	2.943	-26.67	5.058
VH	-16.418	3.414	-35.35	-4.289
`smap_am`	0.147	0.122	0.0	0.675
`soil_moisture`	0.174	0.109	0.0	0.485

Table 2: Descriptive statistics of dataset variables (after outlier removal)

Key Observations:

1. **VV and VH:** Both radar backscatter variables are predominantly negative because backscatter is expressed in decibels (dB), where negative values indicate that reflected energy is weaker than transmitted energy. The high standard deviations (2.9 and 3.4 dB) indicate substantial variability across different surface conditions.

2. **smap_am:** Approximately 19% of values are exactly zero, potentially representing extremely dry conditions or sensor detection thresholds. The mean value of 0.147 indicates relatively low average moisture content.

3. **soil_moisture:** The target variable shows moderate variability with standard deviation of 0.109. The maximum value of 0.485 represents relatively high moisture content.

3.4 Data Quality Assessment

Strengths:

- Complete dataset with zero missing values.
- No duplicate observations.
- All numerical features suitable for regression modeling.
- Reasonable sample size (30,747 observations) for machine learning.
- Balanced feature scales (all within comparable ranges after standardization).

Limitations Identified:

- Only three predictor variables available; important environmental factors (temperature, rainfall, soil type, vegetation density, temporal features) are absent.
- Approximately 19% zero values in `smap_am` may indicate measurement limitations or specific environmental conditions.
- Five extreme outliers detected in target variable (values > 1000) were removed as data anomalies.
- No temporal information available to model time-series dynamics.
- No spatial coordinates to account for geographic variations.

4 Data Preprocessing

4.1 Data Loading and Initial Inspection

The dataset was loaded from CSV format and underwent systematic inspection to assess data quality and characteristics.

Key points:

- Dataset dimensions: 30,747 rows $\times$ 4 columns.
- All features are float64 data type.
- Zero missing values confirmed.
- All variables are continuous numerical features.

4.2 Outlier Detection and Treatment

Exploratory visualization revealed extreme outliers in the target variable `soil_moisture` that required investigation.

Five observations exhibited extreme soil moisture values exceeding 1000 (ranging from 1383.38 to 1396.57), which is physically unrealistic for volumetric soil moisture (valid range: 0–1). These outliers:

- Represent only 0.016% of the dataset (5 out of 30,747 observations).
- Are approximately 3,400 times larger than the mean value.
- Have normal-range values for predictor variables (VV, VH, `smap_am`).
- Strongly suggest data entry errors, unit mismatches, or measurement anomalies.

Treatment Decision These five extreme outliers were removed from the dataset because they:

1. Violate physical constraints of soil moisture measurement.
2. Represent negligible proportion of data (0.016%).
3. Would severely distort regression model training.
4. Cause visualization difficulties by stretching axis scales.
5. Inflate error metrics (RMSE) disproportionately.

After removal, the maximum `soil_moisture` value is 0.485, which is physically realistic.

Justification Statement *During exploratory data analysis, five extreme outlier observations were identified where* ***`soil_moisture`*** *values exceeded 1000 (range: 1383–1397). Since typical volumetric soil moisture values range between 0 and 1, and these outliers represent only 0.016% of observations with otherwise normal predictor values, they were classified as data anomalies and removed to prevent distortion of model training and evaluation metrics. Post-removal, the maximum* ***`soil_moisture`*** *value is 0.485, which is physically realistic.*

4.3 Feature Engineering

No complex feature engineering was performed in this baseline analysis. The focus remained on understanding the predictive capability of the original satellite-derived features. Degree-2 polynomial features (squared terms and pairwise interactions) were later explored and are discussed in the results section.

4.4 Train-Test Split

The dataset was partitioned into training and testing sets:

- **Training Set:** 24,593 samples (80%).
- **Testing Set:** 6,149 samples (20%).
- **Random Seed:** 42.
- **Stratification:** Not applicable (continuous target).

The 80–20 split provides sufficient training data while reserving adequate samples for unbiased performance evaluation on unseen data.

4.5 Feature Scaling

Standardization (z-score normalization) was applied to transform features to zero mean and unit variance:

$$z = \frac{x - \mu}{\sigma}$$

where x is the original feature value, μ is the training set mean, and σ is the training set standard deviation.

Rationale

- Prevents feature dominance across different scales.
- Improves performance of distance-based algorithms and gradient descent.
- Preserves distribution shape better than min–max normalization.

The scaler was fit only on training data and then applied to test data to prevent data leakage.

4.6 Preprocessing Pipeline Summary

The complete preprocessing workflow:

1. Data loading and initial inspection.
2. Missing value verification (none found).
3. Outlier detection and removal (5 extreme values removed).
4. Train-test split (80–20 ratio).

5. Feature standardization (zero mean, unit variance).

6. Data integrity checks passed.

Final dataset:

- Total samples: 30,742 observations.
- Training samples: 24,593.
- Testing samples: 6,149.
- Features: 3 standardized predictors.
- Target: 1 continuous variable (`soil_moisture`).

5 Exploratory Data Analysis

5.1 Distribution Analysis

Univariate distributions were examined using histograms.

VV Distribution

VV backscatter is approximately bell-shaped with mean -9.196 dB and standard deviation 2.943 dB. Values are predominantly between -15 dB and -5 dB, with few extremes. This is consistent with typical C-band backscatter behavior for land surfaces.

VH Distribution

VH backscatter shows a similar approximately normal shape but centered at lower values (mean -16.418 dB) with standard deviation 3.414 dB. This reflects weaker cross-polarized scattering compared to co-polarized VV and greater variability due to vegetation structure.

SMAP Morning Soil Moisture (`smap_am`)

The `smap_am` variable is strongly right-skewed with a large spike at exactly 0.0 (about 19% of observations). Values mostly lie between 0.05 and 0.30, with a long tail up to 0.675. This suggests frequent dry conditions with occasional wet events.

Target Variable (`soil_moisture`)

After removing five extreme anomalies, `soil_moisture` is moderately right-skewed, concentrated between 0.0 and 0.2, with a maximum of 0.485. This reflects typical soil moisture behavior in regions with irregular precipitation.

A $\log(1 + x)$ transformation reduces skewness but does not fully normalize the distribution, indicating inherent non-normality.

5.2 Correlation Analysis

The Pearson correlation matrix for VV, VH, `smap_am`, and `soil_moisture` is:

	VV	**VH**	**smap_am**	**soil_moisture**
VV	1.000	0.888	0.141	-0.041
VH	0.888	1.000	0.278	-0.065
smap_am	0.141	0.278	1.000	0.041
soil_moisture	-0.041	-0.065	0.041	1.000

Table 3: Pearson correlation coefficients between variables

Key observations:

- VV and VH are highly correlated (0.888), indicating multicollinearity.
- Correlations between predictors and target are extremely weak ($|\rho| \leq 0.065$).
- `smap_am` has weak positive correlations with VV and VH.

Heatmaps and pairplots confirm:

- Strong linear relationship between VV and VH.
- No clear linear trends between each predictor and `soil_moisture`.
- No obvious non-linear curvature in bivariate scatterplots.

5.3 Multivariate Visualization

A scatter plot matrix (pairplot) for VV, VH, `smap_am`, and `soil_moisture` shows:

- Tight linear band for VV vs VH.
- Cloud-like scatter with no discernible trends for predictors vs target.
- Uniform scatter consistent with high noise-to-signal ratio.

5.4 Boxplot Analysis and Outlier Quantification

Boxplots for all variables show:

- VV and VH have a few outliers in both tails but within physically reasonable ranges.
- `smap_am` and `soil_moisture` exhibit some upper-tail outliers corresponding to wet conditions.

Outlier counts using the IQR method:

Variable	**Outlier Count**	**Percentage**
VV	1,475	4.8%
VH	1,413	4.6%
`smap_am`	829	2.7%
`soil_moisture`	37	0.12%

Table 4: Outlier frequencies by IQR method (after extreme value removal)

These outliers were retained because they are physically plausible and represent natural environmental variability.

5.5 Summary of EDA Findings

- Clean dataset with no missing values; five extreme outliers removed from the target.
- VV and VH approximately normal; `smap_am` and `soil_moisture` right-skewed.
- Severe multicollinearity between VV and VH; extremely weak predictor-target correlations.
- No strong linear or obvious non-linear relationships visible between predictors and target.

These findings indicate that predicting `soil_moisture` from these three variables alone will be challenging.

6 Model Development

6.1 Baseline Model – Linear Regression

Linear Regression was used as a baseline. After training on scaled features:

- R^2: 0.010.

- RMSE: 0.1088.
- MAE: 0.0899.

Given the target's standard deviation of 0.109, this performance is barely better than predicting the mean for all samples.

Regression coefficients (with standardized features) were small in magnitude and mixed in sign, reflecting weak and unstable relationships.

6.2 Assumption Validation for Linear Regression

Linearity

Residuals vs fitted values showed a random horizontal band around zero with no clear patterns, suggesting the linearity assumption holds reasonably well.

Normality of Residuals

The residual distribution was moderately right-skewed but roughly centered around zero. A Shapiro–Wilk test indicated some deviation from perfect normality, but not enough to explain the poor performance.

Multicollinearity

Variance Inflation Factors (VIF) were:

- VV: ≈ 40.
- VH: ≈ 42.
- `smap_am`: ≈ 2.2.

VV and VH exhibit severe multicollinearity. Removing VH and retraining reduced R^2 to 0.0028, confirming that multicollinearity is not the primary cause of poor performance; the core issue is weak signal.

6.3 Advanced Regression Models

Polynomial Regression (Degree 2)

Adding quadratic and interaction terms:

- R^2: 0.015.
- RMSE: 0.1085.

Only negligible improvement, indicating the absence of strong polynomial relationships.

Ridge Regression

Ridge Regression with $\alpha = 1.0$:

- R^2: 0.010 (same as Ordinary Least Squares).

Random Forest Regressor

Random Forest (100 trees):

- Test R^2: 0.0039.
- 5-fold CV mean R^2: -0.077.

Gradient Boosting Regressor

Gradient Boosting (500 estimators):

- CV mean R^2: 0.011.

Support Vector Regression (Linear Kernel)

SVR with linear kernel:

- Test R^2: 0.0094.
- RMSE: 0.1088.

Support Vector Regression (RBF Kernel)

SVR with RBF kernel:

- Test R^2: 0.0534.
- RMSE: 0.1064.
- MAE: 0.0877.
- 5-fold CV mean R^2: 0.013.

This is a notable improvement relative to linear approaches, though still explaining only about 5% of variance.

6.4 Deep Learning – Neural Network

A fully connected neural network with architecture 64–32–16–1 and ReLU activations was trained with Adam optimizer and MSE loss.

Performance:

- Test R^2: 0.0565.
- RMSE: 0.1062.
- MAE: 0.0872.

Training and validation losses converged without overfitting, but absolute performance remained weak, reinforcing the conclusion of limited predictive signal in the features.

6.5 Model Comparison Summary

Model	**Test R^2**	**RMSE**	**MAE**
Linear Regression	0.010	0.109	0.090
Polynomial Regression (Deg 2)	0.015	0.109	0.090
Ridge Regression	0.010	0.109	0.090
Random Forest	0.004	0.109	0.090
Gradient Boosting	0.011	0.109	0.089
SVR (Linear Kernel)	0.009	0.109	0.090
SVR (RBF Kernel)	**0.053**	**0.106**	**0.088**
Neural Network	**0.056**	**0.106**	**0.087**

Table 5: Performance comparison of all regression models

7 Implementation Details

7.1 Environment and Libraries

The full implementation is done in Python using:

- `pandas`, `numpy` for data manipulation.
- `matplotlib`, `seaborn` for visualization.
- `scikit-learn` for preprocessing, modeling, and evaluation.
- `TensorFlow/Keras` for neural network training.
- `statsmodels` for VIF analysis.

- `joblib` for model persistence.

Random seeds were set (NumPy and TensorFlow) to ensure reproducibility.

7.2 Preprocessing Pipeline

A dedicated preprocessing pipeline:

- Loads the CSV dataset.
- Removes extreme outliers from `soil_moisture`.
- Splits features and target.
- Performs train-test split.
- Applies standardization to the features.

7.3 Model Training and Evaluation Framework

A generalized model evaluation class:

- Accepts both scaled and unscaled feature matrices.
- Trains models and computes R^2, RMSE, and MAE.
- Aggregates and ranks model results.
- Provides optional plots comparing model performance.

7.4 Neural Network Implementation

A Keras Sequential model:

- Input: 3-dimensional standardized feature vector.
- Hidden layers: 64, 32, and 16 neurons with ReLU activations.
- Output: 1 neuron for regression.
- Loss: MSE.
- Optimizer: Adam.

Training:

- 100 epochs.
- Batch size 32.
- Validation split 0.2.

7.5 Model Deployment Artifacts

For deployment:

- The best-performing SVR-RBF model is retrained on training data and saved using `joblib`.
- The fitted `StandardScaler` is saved as well.
- A utility function `predict_soil_moisture(VV, VH, smap_am)` loads both artifacts, scales inputs, and outputs a predicted soil moisture value.

7.6 Requirements

The following versions (or compatible equivalents) are used:

- `pandas 1.5.3`
- `numpy 1.24.3`
- `matplotlib 3.7.1`
- `seaborn 0.12.2`
- `scikit-learn 1.3.0`
- `tensorflow 2.13.0`
- `statsmodels 0.14.0`
- `joblib 1.3.0`

8 Results and Evaluation

8.1 Overall Performance Summary

Across eight regression models, the maximum test R^2 was approximately 0.056 (Neural Network), with SVR-RBF close behind at 0.053. All RMSE values were close to the target's standard deviation (0.109), indicating models only marginally improved over predicting the mean.

8.2 Key Performance Insights

- Only non-linear models (SVR-RBF, Neural Network) extracted slightly more signal, but still explained less than 6% of variance.
- Random Forest exhibited negative mean CV R^2, meaning it performed worse than the mean predictor on validation folds.
- The closeness of RMSE to the baseline (standard deviation of target) mathematically demonstrates weak predictive capability.

8.3 Statistical Significance Analysis

Given target variance $\sigma_y^2 \approx 0.0119$ and best R^2 around 0.056, the explained variance is only about:

$$0.056 \times 0.0119 \approx 0.00067,$$

which is negligible in practice.

8.4 Model Selection Justification

SVR with RBF kernel is selected as the final model due to:

- Its relatively high test R^2 (0.053) with stable cross-validation performance.
- Non-linear modeling capability via RBF kernel.
- Simpler deployment and fewer hyperparameters compared to Neural Networks.

8.5 Error Analysis

Predicted vs actual plots for SVR-RBF show:

- Predictions clustered around a narrow band near the mean, rather than following the identity line.
- Underprediction for high actual soil moisture and overprediction for very low values.

This regression-to-the-mean behavior is consistent with limited predictive information in the features.

9 Discussion

9.1 Interpretation of Results

Multiple lines of evidence indicate that the three predictors (VV, VH, `smap_am`) contain very limited information about the target `soil_moisture` in this dataset:

- Extremely low linear correlations.
- Consistently poor performance across diverse models.
- Assumption checks showing no major statistical violations.
- Minimal gains from polynomial features or regularization.

9.2 Root Cause Analysis

Key reasons for poor performance:

1. **Missing critical environmental variables:** No information on rainfall, temperature, soil properties, vegetation indices, topography, or temporal context.
2. **Scale mismatch:** Sentinel-1 operates at 10–20 m resolution, while SMAP is at ∼36 km, causing spatial averaging and noise.
3. **Physical complexity:** Radar backscatter is influenced by surface roughness, vegetation attenuation, and soil dielectric properties in non-trivial ways.
4. **Low target variance and skewness:** Soil moisture is concentrated in a narrow range with skewed distribution, limiting the variance that can be explained.

9.3 Model Behavior

SVR-RBF and the Neural Network outperform linear methods because they can capture weak non-linear relationships and interactions, even when overall signal is low. Tree-based ensembles underperform due to overfitting noise and the absence of strong, high-contrast splits in feature space.

9.4 Comparison with Literature

Many published studies report R^2 values between 0.5 and 0.85 for soil moisture prediction using Sentinel-1 and SMAP data [1,7,10]. Differences from this study include:

- Use of additional predictors (NDVI, DEM, precipitation).

- More homogeneous study areas.
- Carefully validated in-situ ground truth data.
- Advanced preprocessing (e.g., incidence angle normalization, vegetation correction).

9.5 Scientific Contribution

This study:

- Provides a transparent, end-to-end ML workflow with rigorous diagnostics.
- Demonstrates that radar backscatter alone (with SMAP) may be insufficient in some contexts.
- Highlights the importance of auxiliary environmental data for soil moisture prediction.
- Offers a reproducible baseline against which enhanced models can be compared.

9.6 Limitations

- Only three predictors; no spatial or temporal features.
- No detailed hyperparameter optimization or advanced ensembles (stacking).
- Quality and representativeness of target variable depend on upstream processing.

10 Conclusion

This project implemented a complete soil moisture prediction pipeline using Sentinel-1 SAR VV and VH backscatter, combined with SMAP morning soil moisture estimates. Multiple regression models were rigorously evaluated, including linear methods, tree ensembles, support vector regression, and deep neural networks.

Despite comprehensive modeling efforts, the best models explained only 5–6% of variance in soil moisture, with RMSE values comparable to baseline prediction by the mean. This demonstrates that the available features capture only a very weak signal related to `soil_moisture` for this particular dataset.

The study's main conclusion is that:

> **Accurate soil moisture prediction from satellite data requires integration of additional environmental, temporal, and spatial information beyond raw radar backscatter and coarse-scale SMAP estimates.**

While predictive performance is low, the methodological rigor, honest reporting of negative results, and detailed analysis provide a valuable foundation for future soil moisture modeling work.

11 Future Scope

11.1 Data Enhancement

- Integrate meteorological data: rainfall, temperature, evapotranspiration.
- Include soil properties: texture, organic matter, bulk density.
- Add topographic variables: elevation, slope, aspect, wetness index.
- Use vegetation indices (e.g., NDVI, LAI) from optical sensors.
- Incorporate temporal features and build time-series models.

11.2 Advanced Modeling Techniques

- Stacked ensembles combining SVR-RBF, Gradient Boosting, and Neural Networks.
- Physics-informed ML that leverages radiative transfer and water balance models.
- Sequence models (LSTM, GRU, Transformers) for temporal dynamics.
- Bayesian methods and uncertainty quantification.

11.3 Data Collection and Validation

- Coordinate with field campaigns for high-quality in-situ measurements.
- Use multi-sensor fusion (Sentinel-1, Sentinel-2, SMAP, MODIS).
- Expand spatial and temporal coverage, including multiple years and diverse regions.

11.4 Operational Deployment

- Build real-time prediction pipelines using cloud infrastructure.
- Develop web and mobile applications for farmers and water managers.
- Integrate soil moisture predictions into irrigation control and drought monitoring systems.

11.5 Ethical and Societal Considerations

- Ensure privacy and responsible use of agricultural data.
- Promote equitable access to soil moisture information, especially for smallholder farmers.
- Consider environmental sustainability and water conservation impacts.

12 References

1. HESS. (2022). Exploring the combined use of SMAP and Sentinel-1 data for downscaling soil moisture beyond the 1 km scale. *Hydrology and Earth System Sciences*, 26, 3337–3357. https://hess.copernicus.org/articles/26/3337/2022/
2. NASA JPL. (2015). Soil Moisture Active Passive (SMAP) Mission. *NASA Jet Propulsion Laboratory*. https://smap.jpl.nasa.gov/
3. European Space Agency. (2014). Sentinel-1: ESA's Radar Observatory Mission for GMES Operational Services. *ESA SP-1322/1*. https://sentinel.esa.int/web/sentinel/missions/sentinel-1
4. Entekhabi, D., et al. (2010). The Soil Moisture Active Passive (SMAP) Mission. *Proceedings of the IEEE*, 98(5), 704–716.
5. Bhogapurapu, N., Dey, S., Homayouni, S., Bhattacharya, A., & Rao, Y.S. (2022). Field-scale soil moisture estimation using Sentinel-1 GRD SAR data. *Advances in Space Research*, 70(12), 3845–3858.
6. Torres, R., et al. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment*, 120, 9–24.
7. Paloscia, S., et al. (2013). Soil moisture mapping using Sentinel-1 images: Algorithm and preliminary validation. *Remote Sensing of Environment*, 134, 234–248.
8. Peng, J., et al. (2021). A review of spatial downscaling of satellite remotely sensed soil moisture. *Reviews of Geophysics*, 59(2), e2020RG000723.
9. Chan, S.K., et al. (2016). Assessment of the SMAP passive soil moisture product. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8), 4994–5007.
10. Poojitha, C.L., Deep, S.C., Ramana, R.K.V., Subheesh, A., & Srilakshmi, M. (2024). Prediction of soil moisture using machine learning techniques: A case study of an IoT-based irrigation system. *Irrigation and Drainage*, 73(1), 185–203.

11. Liu, Y., et al. (2023). Machine learning approaches for soil moisture retrieval from synthetic aperture radar: A review. *IEEE Geoscience and Remote Sensing Magazine*, 11(3), 8–35.

12. Fang, K., et al. (2024). A comprehensive study of deep learning for soil moisture prediction. *Hydrology and Earth System Sciences*, 28, 917–943.

13. Kornelsen, K.C., & Coulibaly, P. (2013). Advances in soil moisture retrieval from synthetic aperture radar and hydrological applications. *Journal of Hydrology*, 476, 460–489.

Appendix A: Complete Code Listing

The complete Python codebase for this project is available in the Jupyter Notebook format. All code cells, outputs, visualizations, and explanations are included in the notebook file `Soil-1.ipynb`.

Appendix B: Hardware and Software Environment

Hardware

- GPU: Tesla T4 (Google Colab).
- GPU Memory: 14.4 GB.
- CPU: Intel Xeon (Google Colab standard runtime).
- RAM: 12 GB.

Software

- Python Version: 3.10.12.
- TensorFlow Version: 2.15.0.
- Operating System: Ubuntu 22.04.3 LTS (Colab environment).
- Jupyter Notebook: Google Colab.
- Key Libraries: As listed in Section 7.3 (Requirements).

Appendix C: Data Availability Statement

The dataset used in this study (`t_s1_am_6am.csv`) is derived from publicly available Sentinel-1 SAR data (European Space Agency Copernicus Programme) and NASA SMAP mission products. The specific processed dataset used in this analysis has been provided for academic purposes.

- Sentinel-1 data: https://scihub.copernicus.eu/
- SMAP data: https://nsidc.org/data/smap