Soil Moisture Prediction using ML & Deep Learning

Project Documentation

N. Arun Nenavath | B.Tech – Mathematics and Computing

1. Project Overview

This project predicts soil moisture content from Sentinel-1 SAR radar backscatter signals (VV, VH) and SMAP satellite data using machine learning and deep learning regression models.

Objectives

- Perform Exploratory Data Analysis (EDA)
- Apply feature engineering to capture non-linear relationships
- Train and compare 6 ML/DL models
- Evaluate using RMSE, MAE, and R^2

Dataset

Attribute	Details
Total Samples	30,747 (cleaned: 30,712)
Target Variable	soil_moisture (0–1 range)
Input Features	VV (dB), VH (dB), smap_am
Missing Values	None
Removed Rows	13 duplicates + 22 invalid (soil_moisture > 1)

2. Exploratory Data Analysis (EDA)

2.1 Feature Distributions

Histograms for all four variables in the raw dataset:

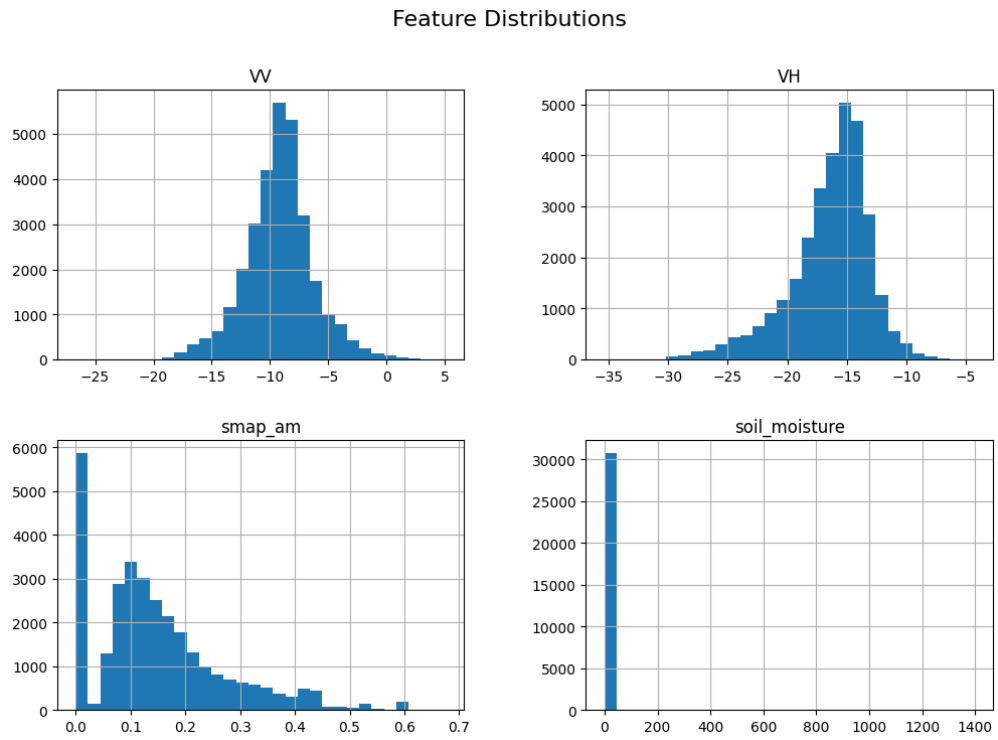

Figure 1: Feature Distributions – VV, VH, smap_am, soil_moisture

- VV: Negatively skewed, centred around −10 dB (normal range for C-band SAR over land)
- VH: Also negatively skewed, centred around −15 dB (cross-pol is weaker than co-pol)
- smap_am: Strongly right-skewed — most observations are dry (values 0.0–0.2)
- soil_moisture: Raw plot shows a spike due to 22 outliers above 1.0 — must be removed before analysis

2.2 Soil Moisture Distribution (After Cleaning)

After removing invalid values, the true distribution becomes visible:

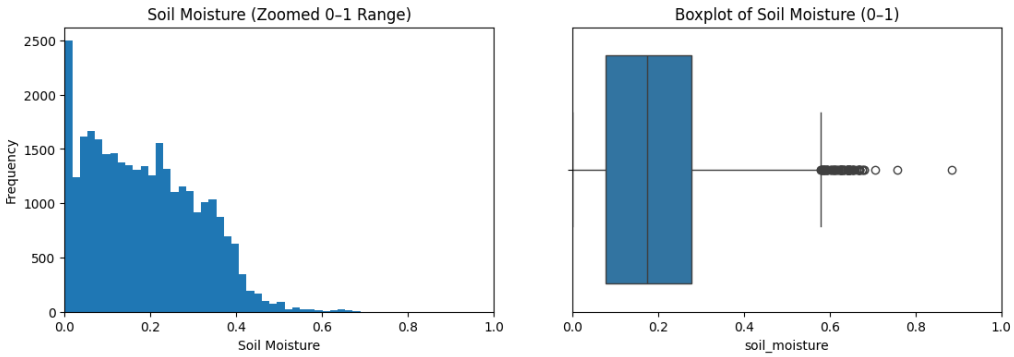

Figure 2: Soil Moisture – Histogram (left) and Boxplot (right)

- Distribution is right-skewed — majority of values fall between 0.0 and 0.3 (dry conditions)
- Median ≈ 0.15, IQR spans 0.05–0.28
- A few valid high-moisture outliers exist in the 0.6–0.9 range (rare rainfall events)
- This imbalance means the model will be more accurate at low moisture and less reliable at high moisture

2.3 Feature vs Soil Moisture – Scatter Plots

Each input feature plotted against the target variable:

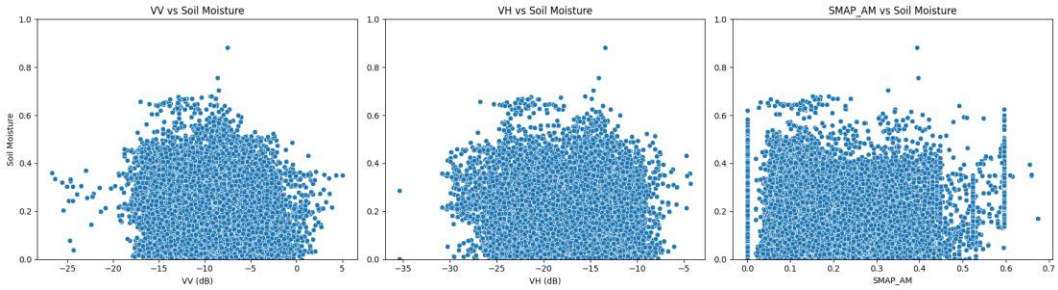

Figure 3: VV, VH, and SMAP_AM vs Soil Moisture

- VV vs Soil Moisture: Completely scattered — no upward or downward trend visible
- VH vs Soil Moisture: Same result — wide, uniform cloud with no pattern
- SMAP_AM vs Soil Moisture: Two vertical clusters visible at 0.0 and 0.6 (likely SMAP quantisation artefacts)
- Conclusion: No single feature shows a linear relationship with soil moisture

2.4 Correlation Analysis

Pearson correlation heatmap showing all pairwise relationships:

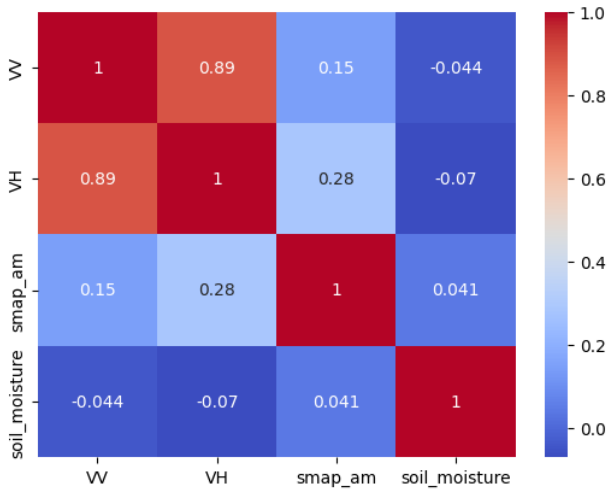

Figure 4: Pearson Correlation Heatmap

Feature	Correlation with soil_moisture	Meaning
VH	−0.070	Very weak negative — barely above noise
VV	−0.044	Very weak negative
smap_am	0.041	Very weak positive
VV ↔ VH	0.89	Very strong — both are SAR backscatter, nearly redundant in raw form

Key takeaway: All feature-target correlations are below 0.07. Linear models will not work. The high VV–VH correlation (0.89) means their ratio and difference carry new, non-redundant information — this directly motivates feature engineering.

3. Feature Engineering

Since raw features have near-zero linear correlation with soil moisture, new features were created to expose non-linear patterns:

Transformation	New Features Created	Why
dB → Linear	VV_lin, VH_lin	Converts logarithmic scale to linear for better model learning
Difference / Ratio	VV−VH, VV/VH, VH/VV	Extracts contrast between the two polarisations
Interaction	VV×smap, VH×smap	Captures joint effect of SAR + SMAP
Polynomial	VV^2, VH^2, $smap^2$	Models non-linear / accelerating relationships
Log Transform	log(1+smap)	Compresses right-skewed SMAP distribution

Final 8 selected features: VV_lin, VH_lin, smap_am, VV_minus_VH, VV_VH_ratio, VV_smap_interaction, VH_smap_interaction, smap_sq

4. Model Results

Before Tuning

Model	RMSE	MAE	R^2
Gradient Boosting	0.01477	0.10186	0.072
SVR (RBF)	0.01503	0.10271	0.056
Random Forest	0.01546	0.10263	0.029
XGBoost	0.01678	0.10601	−0.054
KNN	0.01681	0.10612	−0.056

After Hyperparameter Tuning

Model	RMSE	MAE	R^2	Best Params
Random Forest ✓	0.01458	0.10076	0.084	n_estimators=600, max_depth=15
Gradient Boosting	0.01467	0.10130	0.079	n_estimators=200, lr=0.1
XGBoost	0.01474	0.10170	0.074	n_estimators=300, max_depth=3

Best Model: Tuned Random Forest — lowest RMSE (0.01458) and highest R^2 (0.084)

5. Residual Analysis

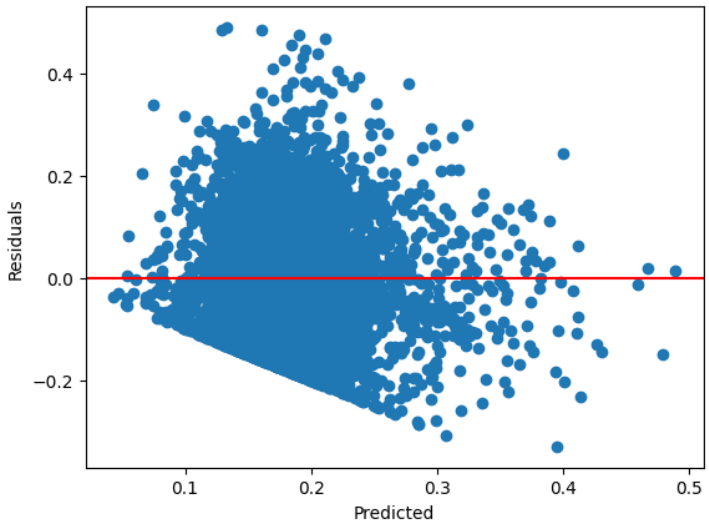

Figure 5: Residual Plot – Predicted vs Residuals (Random Forest)

- Residuals are centred around zero — no systematic bias
- Funnel shape (wider at higher predictions) indicates heteroscedasticity
- Model is more accurate for low-moisture values (majority of training data) and less reliable at higher moisture

6. Deep Learning Model (ANN)

Architecture

Input → Dense(128) → Dense(64) → Dense(32) → Dense(16) → Output | Activation: ReLU | Optimizer: Adam (lr=0.001) | Early Stopping: patience=10

Metric	ANN	Best ML (Tuned RF)
RMSE	0.01496	0.01458
MAE	0.10266	0.10076
R^2	0.061	0.084

Training Curve

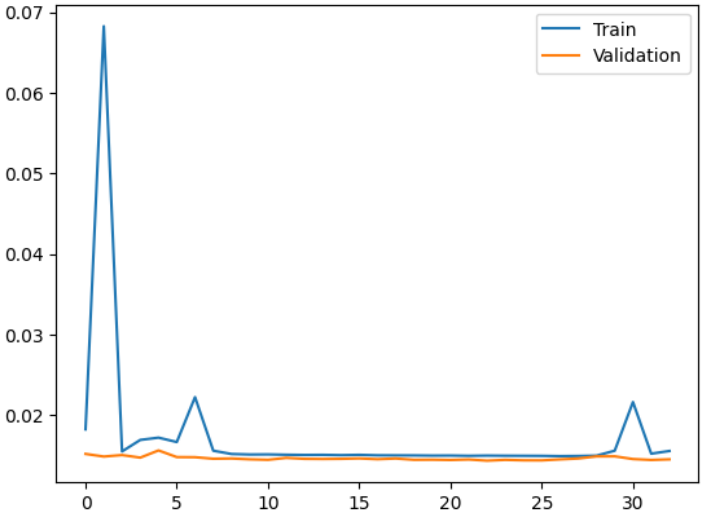

Figure 6: ANN Training and Validation Loss

- Training and validation loss converge — no overfitting
- Loss stabilises early, indicating the model has extracted all available signal from the 8 features
- Deep learning did not beat Random Forest — tree ensembles perform better on small tabular datasets

7. Key Findings

Finding	Conclusion
Feature-target correlations < 0.07	Linear models will fail — non-linear approach required
VV and VH correlation = 0.89	Ratio and difference features add genuinely new information
Feature engineering improved R^2 from ~0 to 0.084	Domain-informed features are critical for this problem
Random Forest outperformed all models	Tree ensembles work best on small structured datasets
ANN underperformed Random Forest	Deep learning needs richer features to compete here
Heteroscedasticity in residuals	Model less reliable at high moisture due to data imbalance

8. Possible Improvements

- Add temporal/seasonal features
- Include geolocation and land cover type
- Ensemble stacking across RF, GB, and XGBoost
- Use TabNet or attention-based models for tabular DL
- Oversample high-moisture events to reduce heteroscedasticity

9. Resume Description

Developed a regression pipeline to predict soil moisture from Sentinel-1 SAR radar features (VV, VH) and SMAP satellite data. Performed EDA on 30,747 samples, applied feature engineering (dB conversion, polarisation ratios, interaction terms), and benchmarked six ML/DL models with hyperparameter tuning. Achieved best R^2 of 0.084 using tuned Random Forest, outperforming Gradient Boosting, XGBoost, SVR, KNN, and a 4-layer ANN.