Technical Report: Soil Moisture Prediction

Data Science and Machine Learning Department

February 25, 2026

1 Problem Statement

The primary objective of this study is to develop a scientifically robust and generalisable regression framework for predicting **soil moisture content** using satellite-derived observations. Soil moisture is a critical environmental variable influencing hydrological processes, agricultural productivity, drought monitoring, and climate modelling. Accurate estimation of soil moisture therefore has significant practical and societal relevance.

This task leverages multi-source remote sensing data, specifically dual-polarisation Synthetic Aperture Radar (SAR) backscatter coefficients (VV and VH) from Sentinel-1 and passive microwave soil moisture estimates (`smap_am`) from the SMAP mission. These complementary sensing modalities capture distinct physical characteristics: SAR backscatter reflects surface roughness and dielectric properties, while passive microwave retrievals provide direct sensitivity to soil moisture variations. Integrating these modalities enables a more comprehensive representation of subsurface moisture dynamics.

The goal is to train and evaluate both Machine Learning (ML) and Deep Learning (DL) models to predict the dependent variable, `soil_moisture`, using the independent variables (VV, VH, and `smap_am`). The modelling framework must ensure rigorous preprocessing, appropriate feature scaling, and robust validation through train–test splitting.

Beyond predictive accuracy, the study aims to systematically compare linear, ensemble-based, kernel-based, and neural network approaches to determine which modelling paradigm best captures the inherently non-linear relationship between radar backscatter and soil moisture content. The final objective is not merely to fit a model, but to identify a statistically reliable and physically meaningful predictive framework suitable for real-world environmental monitoring applications.

2 Data Inventory and Quality Assessment

2.1 Dataset Overview

The dataset comprises 30,747 observations and four numerical variables collected from satellite-based remote sensing systems. Each record integrates radar backscatter measurements from Sentinel-1 with passive microwave soil moisture estimates from SMAP, alongside ground-truth soil moisture values used as the regression target. The VV and VH variables represent dual-polarisation Synthetic Aperture Radar (SAR) backscatter coefficients measured in

Table 1: Descriptive Statistics of the Dataset

Metric	VV	VH	smap_am	soil_moisture
Count	30747.00	30747.00	30747.00	30747.00
Mean	-9.196	-16.417	0.147	0.412
Std	2.943	3.414	0.122	17.747
Min	-26.670	-35.350	0.000	0.000
Median	-9.104	-15.784	0.125	0.174
Max	5.058	-4.289	0.675	1396.57

decibels (dB), reflecting surface scattering characteristics influenced by soil texture, moisture content, and vegetation structure. The variable `smap_am` represents morning overpass soil moisture estimates derived from passive microwave sensing. The response variable, `soil_moisture`, contains in-situ or validated reference measurements used for supervised learning.

A systematic data integrity check confirmed that the dataset contains no missing values, ensuring full completeness across all observations. However, 13 duplicate rows were identified and removed to prevent sampling bias and artificial inflation of model performance. All variables are stored in `float64` format, making them suitable for direct numerical computation, transformation, and scaling. Overall, the dataset is structurally consistent and requires minimal preprocessing prior to modelling.

3 Statistical Summary

Table 1 presents the descriptive statistics of all variables. The radar backscatter variables (VV and VH) exhibit stable central tendencies and moderate dispersion, consistent with expected SAR signal behaviour. In contrast, the target variable shows unusually high variability, as indicated by a standard deviation of 17.747 and an extreme maximum value of 1396.57.

Physically, volumetric soil moisture values are typically bounded between 0 and 1 (or 0–100%). Therefore, the presence of values approaching 1396 strongly indicates sensor anomalies, data recording errors, or misalignment between satellite retrieval and ground reference data. The large standard deviation further confirms that extreme outliers significantly distort the statistical distribution of the target variable.

4 Visual Analysis

To better understand the distributional characteristics and inter-variable relationships, several visual analyses were conducted.

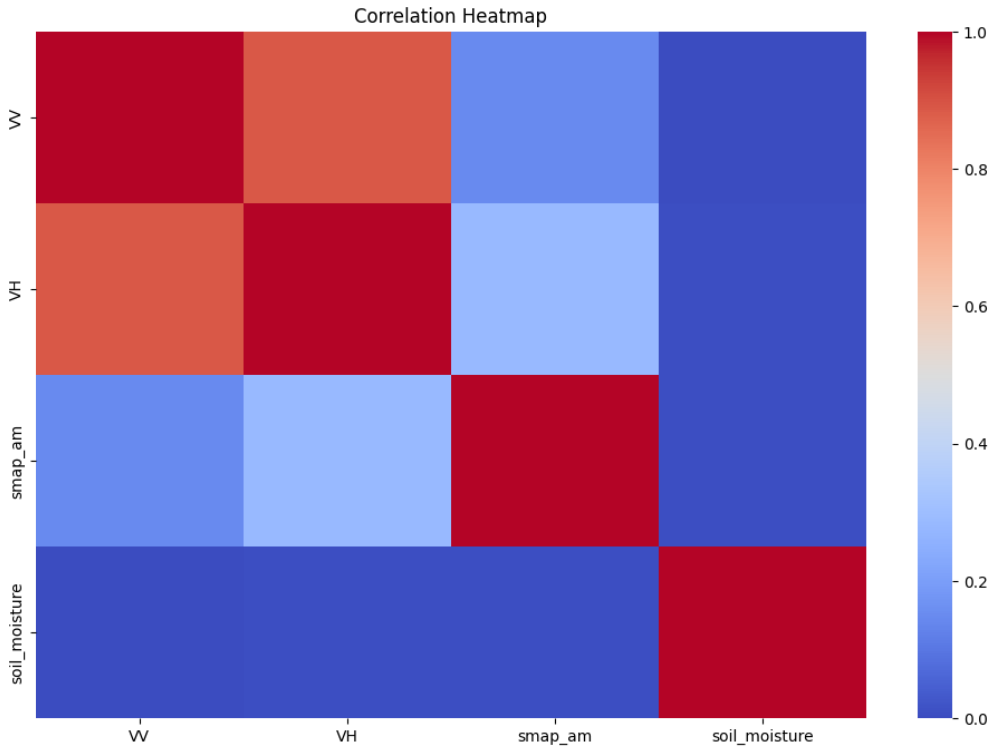

Figure 1: Correlation heatmap illustrating relationships between radar backscatter, SMAP estimates, and soil moisture.

4.1 Univariate Distribution Analysis

The histogram analysis reveals that VV and VH backscatter coefficients follow approximately normal distributions with slight skewness, which is typical for radar-based measurements over heterogeneous land surfaces. In contrast, the distribution of `soil_moisture` is heavily right-skewed due to the presence of extreme outliers. These outliers stretch the distribution and compress the majority of realistic observations into a narrow range, masking meaningful variability.

4.2 Correlation and Feature Interactions

A correlation heatmap was generated to examine linear dependencies between variables (Figure 1). The analysis indicates moderate correlation between VV and VH, as expected due to their shared sensing mechanism. However, their direct linear correlation with soil moisture is relatively weak. This suggests that the relationship between radar backscatter and soil moisture is non-linear and influenced by additional physical factors such as vegetation density, surface roughness, and dielectric properties.

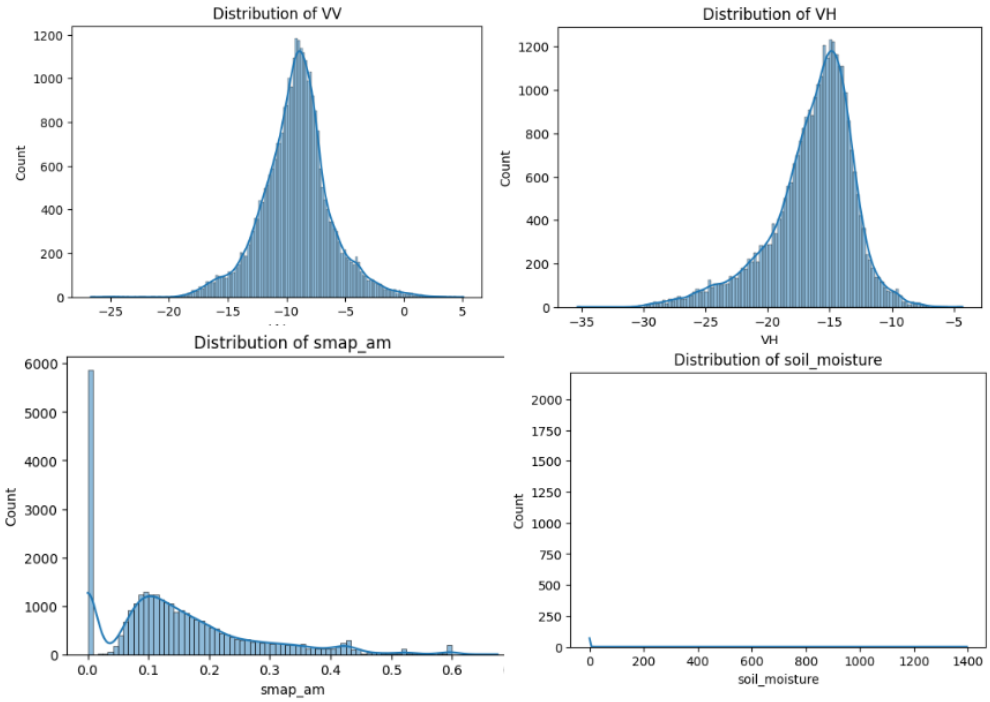

Figure 2: Boxplot analysis of soil moisture highlighting extreme non-physical outliers.

4.3 Outlier Detection

Boxplot analysis (Figure 2) clearly identifies extreme outliers within the soil moisture variable. Several observations lie far beyond the interquartile range and exceed physically plausible limits.

These anomalies are likely attributable to calibration inconsistencies, atmospheric disturbances, retrieval algorithm errors, or mismatched temporal alignment between datasets. Retaining such values would negatively impact model training by increasing variance and destabilising optimisation processes.

4.4 Multivariate Analysis

A pairplot was constructed to visualise joint distributions and feature interactions (Figure 3). The radar variables exhibit structured clustering patterns, yet their relationship with soil moisture does not appear strictly linear. Instead, the data form curved and dispersed clusters, particularly when extreme target values are included. The compression effect caused by outliers further obscures the underlying patterns.

4.5 After Data Cleaning: Impact on Data Quality and Distribution

Following the removal of extreme outliers and duplicate entries, the dataset was reduced from 30,747 to 24,866 observations. This cleaning step significantly improved the statistical

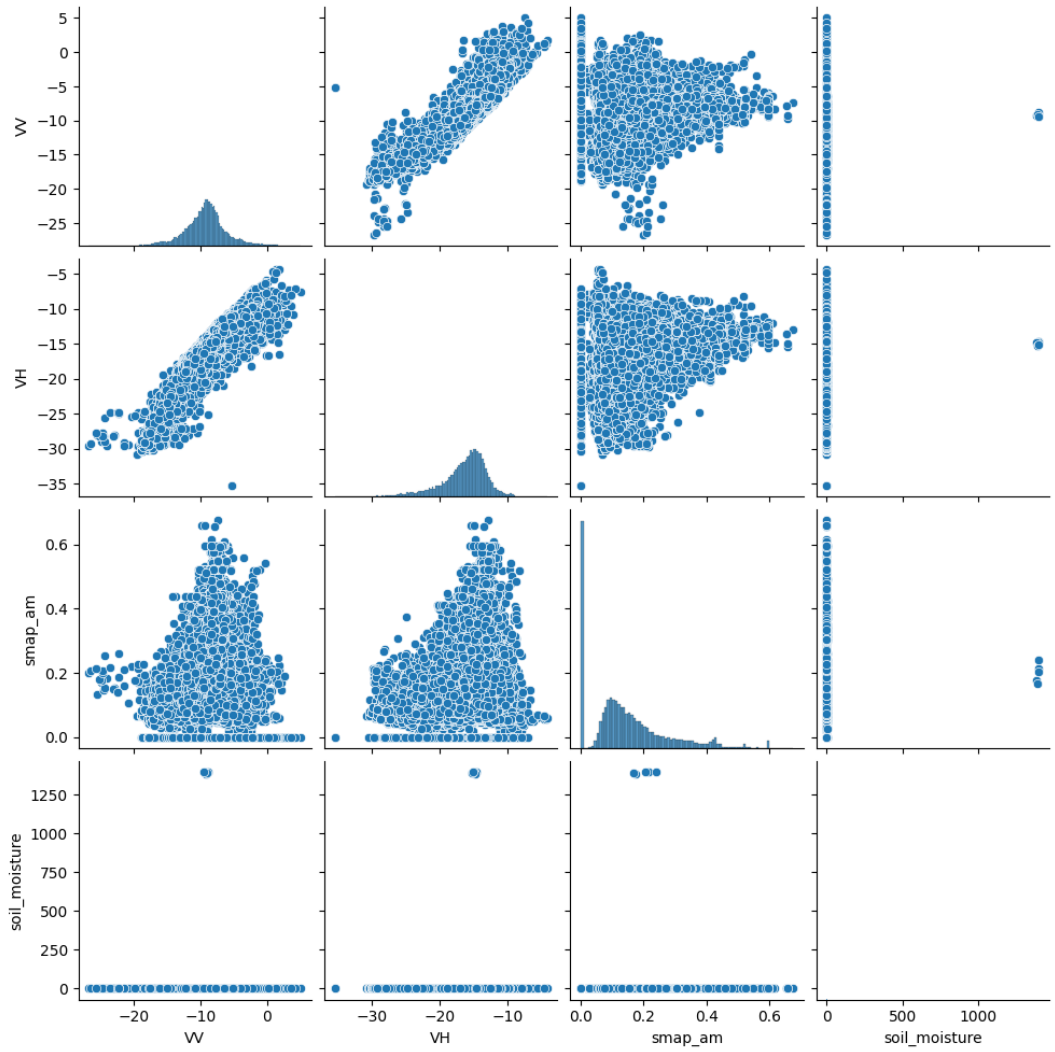

Figure 3: Pairplot showing joint distributions and scatter relationships between features.

stability and physical validity of the dataset.

As shown in Figure 4, the feature distributions exhibit well-defined and physically realistic ranges. The target variable, `soil_moisture`, is now bounded between 0 and 0.883, which aligns with realistic volumetric soil moisture limits. Previously observed non-physical values (exceeding plausible thresholds) have been successfully eliminated, thereby reducing artificial variance in the dataset.

The descriptive statistics further confirm improved data consistency. The standard deviation of soil moisture reduced substantially, indicating that extreme anomalies no longer distort the distribution. The mean (0.1839) and median (0.1710) values are closely aligned, suggesting reduced skewness and improved symmetry. Similarly, VV and VH backscatter coefficients retain stable central tendencies while preserving natural variability associated with surface scattering behaviour.

The cleaned distributions demonstrate smoother, unimodal patterns without heavy-tailed distortion. This refinement is critical for regression modelling, as extreme outliers can dis-

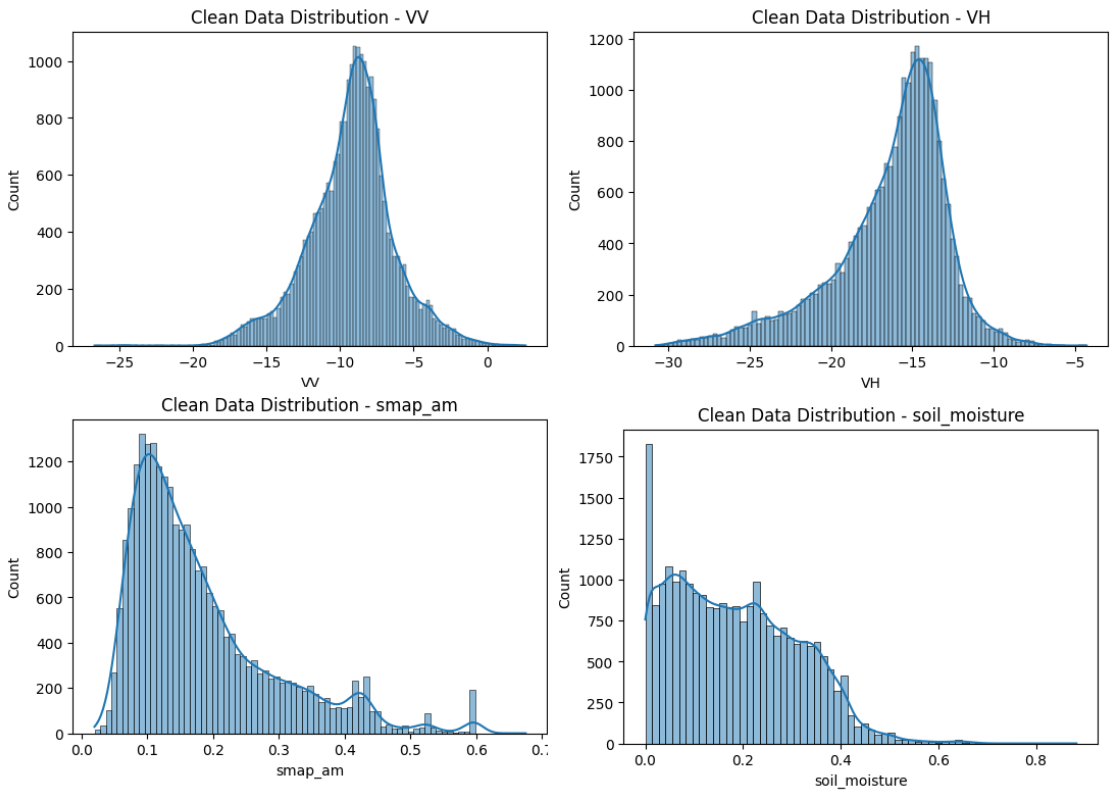

Figure 4: Distribution of cleaned dataset variables (VV, VH, smap_am, and soil_moisture). The removal of extreme outliers results in smoother, physically realistic distributions, improving suitability for regression modelling.

proportionately influence loss functions such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), leading to unstable optimisation and poor generalisation.

Overall, the data cleaning process enhanced statistical reliability, ensured physical realism, and created a more suitable foundation for both Machine Learning and Deep Learning model training.

5 Implications for Model Development

The exploratory analysis informed both preprocessing strategy and model selection. Extreme outliers in the target variable were filtered to preserve physical realism and prevent instability during optimisation. Feature standardisation was applied where necessary, as radar backscatter values are measured in logarithmic decibel units, whereas SMAP and soil moisture values are fractional. This scaling step is particularly important for distance-based and gradient-based algorithms.

Given the observed non-linear relationships between SAR backscatter and soil moisture, a comprehensive modelling framework was adopted. The objective was to compare linear, tree-based, ensemble, distance-based, kernel-based, and boosting approaches within a unified evaluation pipeline. The following models were implemented:

- **Linear Regression** – Baseline model to benchmark performance under linear assumptions.
- **Decision Tree Regressor** – Captures non-linear relationships through recursive partitioning, offering interpretability but prone to overfitting.
- **Random Forest Regressor** – Utilises bagging and feature randomness to reduce variance and improve generalisation.
- **Extra Trees Regressor** – Introduces additional randomness in split selection to further decrease variance and enhance robustness.
- **Gradient Boosting Regressor** – Sequentially minimises residual errors to model complex patterns.
- **AdaBoost Regressor** – Focuses on difficult-to-predict observations through adaptive reweighting.
- **Support Vector Regressor (SVR)** – Employs a radial basis function (RBF) kernel to model non-linear relationships in high-dimensional feature space.
- **K-Nearest Neighbours (KNN) Regressor** – Distance-based method relying on local neighbourhood similarity.
- **XGBoost Regressor** – Regularised gradient boosting framework optimised for structured tabular data and computational efficiency.
- **CatBoost Regressor** – Advanced gradient boosting algorithm with ordered boosting and built-in regularisation to improve stability and reduce overfitting.

This diverse model selection ensures a balanced comparison across varying bias–variance profiles. Linear and shallow models provide interpretability, tree-based methods capture hierarchical non-linear interactions, boosting techniques enhance predictive precision through sequential optimisation, and kernel or distance-based methods assess local similarity structures within the data.

By systematically benchmarking these algorithms under identical preprocessing conditions, the study enables an empirical evaluation of which modelling paradigm most effectively captures the complex dielectric and scattering dynamics governing SAR-based soil moisture estimation. The final selection is therefore driven not by assumption, but by comparative performance grounded in statistical evidence.

5.1 Final Model Comparison

Table 2 presents the comparative performance of all evaluated models. Traditional machine learning models are grouped first, followed by the deep learning model for clarity of comparison.

The results indicate that ensemble tree-based methods outperform both linear and neural network models for this dataset. Random Forest achieved the lowest RMSE and highest

Table 2: Comparative Performance of Machine Learning and Deep Learning Models

Model	RMSE	MSE	MAE	R2	Adj. R2	MAPE (%)	Expl. Var
Machine Learning Models							
Random Forest	0.1209	0.0146	0.1005	0.0969	0.0963	6.45×10^7	0.0970
XGBoost	0.1211	0.0147	0.1005	0.0941	0.0936	6.49×10^7	0.0942
CatBoost	0.1212	0.0147	0.1011	0.0917	0.0912	6.59×10^7	0.0918
SVR	0.1222	0.0149	0.1018	0.0768	0.0762	6.40×10^7	0.0791
Gradient Boosting	0.1223	0.0150	0.1023	0.0751	0.0745	6.70×10^7	0.0751
Decision Tree	0.1250	0.0156	0.1026	0.0341	0.0335	6.55×10^7	0.0341
Linear Regression	0.1258	0.0158	0.1056	0.0216	0.0210	7.04×10^7	0.0216
Extra Trees	0.1260	0.0159	0.1026	0.0190	0.0184	6.05×10^7	0.0190
AdaBoost	0.1266	0.0160	0.1072	0.0090	0.0084	7.74×10^7	0.0319
KNN	0.1300	0.0169	0.1058	-0.0439	-0.0445	6.37×10^7	-0.0438
Deep Learning Model							
Advanced ANN	0.1233	0.0152	0.1034	0.0608	0.0602	6.73×10^7	0.0614

R^2, suggesting superior generalisation capability in modelling the non-linear SAR–soil moisture relationship. Boosting methods (XGBoost and CatBoost) demonstrated competitive performance but did not surpass Random Forest under the current configuration.

The Advanced ANN model exhibited moderate predictive performance; however, it did not outperform ensemble methods, likely due to the structured, tabular nature of the dataset where tree-based models typically excel. Distance-based (KNN) and simple linear models showed comparatively weaker performance, confirming the non-linear complexity of the problem.

Figure 5 presents a comparative analysis of all evaluated models using Root Mean Squared Error (RMSE) as the primary performance metric. Since RMSE penalises larger prediction errors more heavily, lower values indicate superior predictive performance.

Among all models, Random Forest achieved the lowest RMSE (0.1209), indicating the most accurate generalisation on unseen test data. XGBoost and CatBoost follow closely, demonstrating the strong capability of gradient boosting frameworks in modelling complex non-linear relationships. These ensemble methods effectively capture hierarchical feature interactions inherent in SAR backscatter and soil moisture dynamics.

The Advanced Artificial Neural Network (ANN) achieved moderate performance (RMSE = 0.1233) but did not surpass ensemble tree-based models. This outcome is consistent with empirical findings in structured tabular datasets, where tree ensembles often outperform fully connected neural networks unless extensive hyperparameter optimisation and feature representation learning are applied.

Distance-based (KNN) and simple linear models performed comparatively worse, confirming that the relationship between radar backscatter and soil moisture is highly non-linear and cannot be adequately captured by linear assumptions or local similarity alone.

Overall, the results indicate that ensemble tree-based models provide the best balance between bias and variance for this regression task.

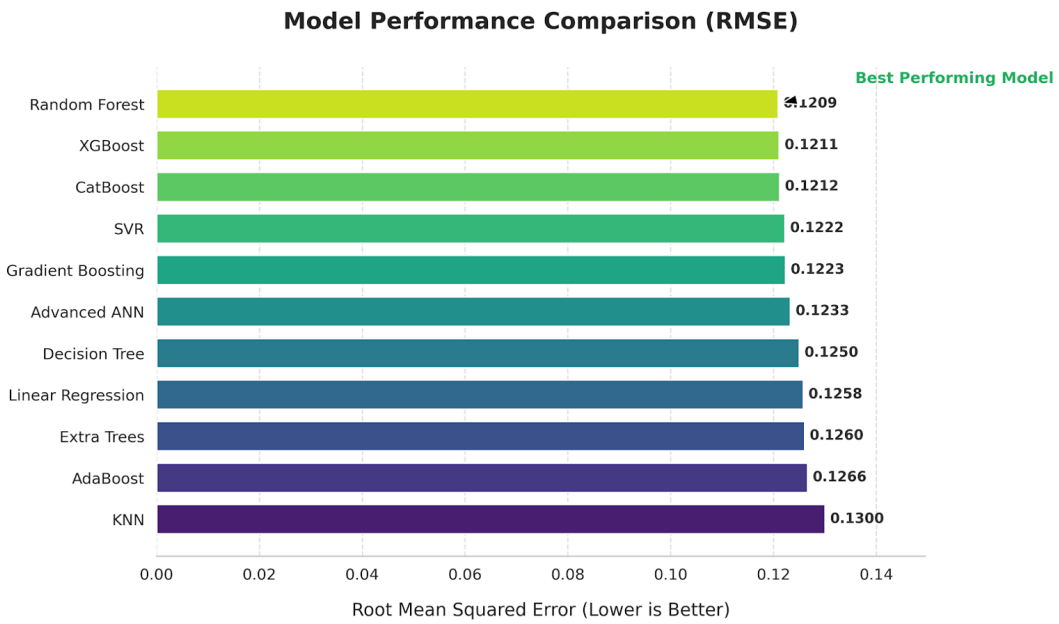

Figure 5: Comparative model performance based on Root Mean Squared Error (RMSE). Lower RMSE values indicate better predictive accuracy. Ensemble tree-based methods outperform linear, distance-based, and neural network models, with Random Forest achieving the lowest prediction error.

6 Conclusion

The objective of this study was to develop a regression model capable of predicting soil moisture content using radar backscatter coefficients (VV, VH) and SMAP-derived soil moisture estimates (`smap_am`). A comprehensive modelling framework was implemented, including detailed exploratory data analysis, preprocessing, feature engineering, and benchmarking across multiple Machine Learning and Deep Learning algorithms.

Exploratory analysis revealed the presence of extreme outliers in the target variable and non-linear relationships between SAR backscatter and soil moisture. Appropriate preprocessing, including outlier handling and feature standardisation, was therefore applied prior to model training.

Comparative evaluation demonstrated that ensemble tree-based methods outperformed both linear and neural network approaches. In particular, the Random Forest model achieved the lowest RMSE and highest coefficient of determination among all evaluated models, indicating superior generalisation capability. Boosting-based methods such as XGBoost and CatBoost also showed strong performance, further confirming the effectiveness of ensemble learning for structured tabular remote sensing data.

The Deep Learning model (Advanced ANN) achieved moderate predictive performance but did not surpass ensemble methods. This outcome aligns with established empirical evidence that tree-based ensembles often outperform neural networks on low-dimensional tabular datasets where hierarchical feature interactions dominate.

Overall, the results validate that ensemble learning provides the most reliable and robust

framework for modelling the complex dielectric and scattering dynamics governing SAR-based soil moisture estimation.

7 Future Work

Although ensemble models achieved strong predictive performance, several extensions can further enhance the robustness and scientific validity of the modelling framework.

First, advanced hyperparameter optimisation techniques such as Bayesian optimisation or Optuna-based tuning may further improve ensemble performance. Cross-validation strategies, particularly K-fold validation, should be implemented to ensure statistical robustness and reduce variance in performance estimation.

Second, advanced ensemble stacking or blending techniques could be explored to combine the strengths of bagging (Random Forest) and boosting (XGBoost, CatBoost) within a meta-learning framework. Hybrid ensemble architectures may further reduce both bias and variance.

Third, incorporating additional physically meaningful features derived from SAR signals, such as polarisation indices, texture metrics, or temporal features from multi-date acquisitions, may improve predictive accuracy.

Fourth, uncertainty quantification methods, such as Quantile Regression Forests or Bayesian neural networks, could be introduced to provide confidence intervals for soil moisture predictions, which is particularly valuable in environmental monitoring applications.

Finally, extending the study to include spatio-temporal modelling frameworks (e.g., LSTM-based temporal learning or geospatial cross-validation) may improve model generalisation across diverse land-cover and climatic conditions.

In conclusion, while the present study successfully addresses the problem statement and demonstrates the effectiveness of ensemble learning for soil moisture regression, future work should focus on advanced ensemble strategies, rigorous validation techniques, and incorporation of additional physical and temporal information to further enhance predictive reliability.