

Explainable Machine Learning Framework for Soil Moisture Stress Assessment and Irrigation Decision Support

Esha Bhausaheb Choudhari¹ and Ms. Asha Mane²

¹Student, Masters of Computer Applications

²Assistant Professor, Masters of Computer Applications

Sadhu Vaswani Institute of Management Studies for Girls, Pune
Savitribai Phule Pune University, Pune, Maharashtra, India.

Abstract: *Water management is one of the vital activities in agriculture, and soil moisture is one of the significant parameters in irrigation planning. The research presents an Explainable Machine Learning Framework for Soil Moisture Stress Assessment and Irrigation Decision Support, using data associated with Pune, Maharashtra. Historical data of soil-moisture, rainfall, and seasonal factors were used in the experiment to predict the soil moisture of the next day. Comparative analysis using Linear Regression, Decision Tree, and Random Forest was performed using chronological splitting, and all three models were assessed using MAE, RMSE, and R^2 . Linear Regression algorithm applied on combined historical soil-moisture, rainfall-history, and seasonal factors recorded the best result with RMSE equals to 1.7977 and R^2 of 0.7819. The analysis using SHAP highlighted the importance of current soil moisture, whereas RMSI categorized the prediction results of the soil moisture for relative irrigation planning and decision-making support.*

Keywords: *Soil Moisture, Machine Learning, Irrigation Decision Support, Explainable AI, SHAP, Relative Moisture Stress Index, Rainfall, Pune, Maharashtra*

I. INTRODUCTION

Agriculture is dependent on water, whose availability in the soil varies according to rainfall, weather, seasons, and previous soil conditions. This study proposes a machine-learning framework to predict the next-day soil-moisture using current and past soil-moisture, rainfall, and seasonal features. The framework evaluates multiple models, comparing their performance using MAE, RMSE, and R^2 , and provides model-interpretation using SHAP. In addition, a Relative Moisture Stress Index (RMSI) is used to interpret the predictions according to the distribution of the training data. Using data from Pune, Maharashtra, the framework combines prediction, model interpretation, and relative stress estimation as an aid in irrigation planning. However, it does not estimate the amount of water required due to the lack of information about crop types, growth stages, soil types, and other factors in the training data.

Objectives of the Research

The objectives of the research are:

- Investigate the application of machine-learning methods for the purpose of soil-moisture assessment and irrigation decision-making.
- Collect and prepare the necessary data on soil moisture, rainfall, weather conditions, and other relevant parameters for the selected study region.
- Construct and compare multiple machine-learning models using the collected data on soil moisture, rainfall, and other factors, as well as the seasonal characteristics of the soil.

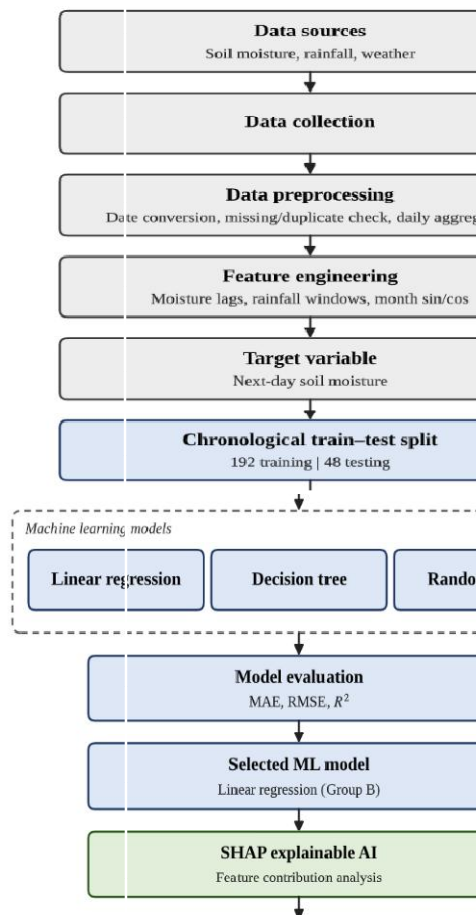


- Evaluate the performance of the developed models using the mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2) metrics, and apply the SHAP-based explainable AI methodology to interpret the results.
- Develop a relative moisture stress index (RMSI) that can be used to estimate the soil-moisture conditions and provide approximate guidance on the corresponding irrigation decisions.

II. RESEARCH METHODOLOGY

The methodology of this research aims at developing an explainable machine-learning framework for the task of next-day soil-moisture prediction and irrigation decision-support. The paper utilizes secondary environmental and agricultural data associated with Pune, Maharashtra. The methodology involves steps of data collection, preprocessing, feature engineering, chronological modeling, model evaluation, Explainable AI (using SHAP), and Relative Moisture Stress Index calculation. The whole methodology follows a time-based framework where the observations of the past are utilized to predict the condition of the soil-moisture for the following day. The predicted value is then explained using SHAP and interpreted using RMSI to provide relative moisture-stress information.

Methodology Diagram



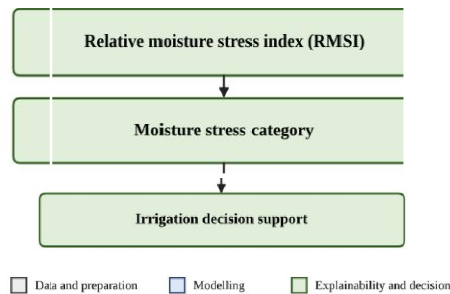


Figure 1: Proposed explainable irrigation decision-support framework

III. LITERATURE REVIEW

Machine learning has been used extensively to predict soil-moisture using the historical values of soil-moisture, rainfall, weather, and environmental variables. Several regression models such as Linear Regression, SVM, Decision Tree, Random Forest, and LSTM are commonly applied for the task while the XAI technique such as SHAP explain the model behavior. However, most of the previous works focus on developing advanced deep-learning models, automated irrigation systems, or crop-specific predictions. This work, therefore, predicts the next-day soil-moisture using historical soil-moisture, rainfall, and seasonal features, compare the performance of various machine learning models, employ the SHAP framework to interpret model predictions, and finally use the Relative Maximum Stress Index (RMSI) to provide relative moisture-stress information as support for decision-making in irrigation.

Research Gap

The literature suggests that machine learning is widely adopted for the purpose of soil-moisture prediction and irrigation management, while the temporal features and XAI are gaining increasing attention. Considering the reviewed literature, the present research focuses on a relatively simple and interpretable integrated workflow that incorporates historical soil-moisture conditions, rainfall history, seasonal features, SHAP-based explanation, and relative moisture-stress interpretation for the purpose of next-day soil-moisture prediction and irrigation decision support in the selected study context. As such, the present research addresses the need for a simple, interpretable, and data-driven framework that integrates soil-moisture prediction with feature-level explanation and relative moisture-stress assessment for the purpose of irrigation decision support.

Research Design

The research will adopt a quantitative, empirical, and experimental research design using secondary environmental and agricultural data related to Pune, Maharashtra, to examine whether historical soil-moisture, rainfall, and seasonal information can predict next-day soil moisture. The study will involve data collection, preprocessing, feature engineering, chronological train-test division, and development and evaluation of machine-learning models using MAE, RMSE, and R^2 , with the persistence method used as a benchmark. Explainable AI using SHAP will be applied to interpret feature contributions, while the Relative Moisture Stress Index (RMSI) will provide relative moisture-stress categories for irrigation purposes. Overall, the research follows data collection → preprocessing → feature engineering → chronological model development → model evaluation → SHAP explanation → RMSI interpretation → irrigation decision support, enabling evaluation of machine-learning methods for soil-moisture prediction and irrigation

Data Collection and Feature Engineering

The study makes use of secondary soil, rainfall, weather, crop-related, and irrigation-related datasets. In particular, the Maharashtra 2020 soil-moisture database is used as the main experimental dataset, while the Pune and Maharashtra rainfall and weather datasets provide additional environmental and seasonal context to the analysis. The soil-moisture



data used in the study were collected at 15-cm depth. Following data cleaning and feature engineering steps, the resulting dataset comprised 240 instances and 24 features, spanning from January 8, 2020, to September 22, 2020. Features were engineered to reflect temporal patterns that may be used to predict subsequent soil-moisture measurements. Three feature sets were defined for the study – Group A comprised features with respect to the current value of the soil's moisture level. Group B was similar to Group A but included additional features related to the lagged (previous) values of the same measurement, current and past rainfall measurements, and seasonal patterns. Finally, Group C was like Group B but also incorporated the current weather conditions. Lagged features reflected prior values of the target variable, while rainfall features incorporated the lagged (previous) value of daily rainfall as well as rolling averages over the previous three- and seven-day rainfall intervals. Seasonality features were engineered using the month as a sinusoidal function. To ensure Temporal consistency and avoid data leakage, rainfall features were lagged, only using prior information available before the day of prediction.

Data Preprocessing

There were several essential preprocessing steps taken before building the model to enhance the quality of data and prevent presenting irrelevant information to the models. The data sets were loaded and explored using Python and Pandas, and the following key preprocessing steps were taken. Check the data set structure, look for missing values, check for duplicates, convert the information about dates to a uniform format, sort the observations chronologically, and align soil-moisture, rainfall, and weather information according to the date. The target variable was created as the next-day soil-moisture value. This allows the model to learn from the information available up to that day without using future information. After preprocessing and features engineering, the final modelling data set had 240 rows and no missing values.

Model Training and Machine Learning Models

The study used a chronological train-test split because the data was a time series. There was no random split to prevent future observations from contaminating the training set. The final dataset was split into 192 train observations from 8 January to 5 August 2020 and 48 test observations from 6 August to 22 September 2020.

Three machine-learning regression models evaluated were Linear Regression, Decision Tree Regression, and Random Forest Regression. The study also adopted a persistence approach as a benchmark for comparison. In the persistence method, the current value of the soil-moisture is used to predict the next-day value. Based on the evaluation results, Linear Regression

with Group B features was selected as the final machine-learning model because it had the lowest RMSE and highest R^2 of all the trained machine-learning models.

Model Evaluation

The models were evaluated using three regression performance measures:

Mean Absolute Error (MAE) – indicates average absolute difference between actual and predicted soil-moisture values.

Root Mean Squared Error (RMSE) – highlights large prediction errors.

R^2 – represents the proportion of variance in the target variable explained by the model.

The selected Linear Regression model using Group B features achieved:

MAE = 1.3630, RMSE = 1.7977, R^2 = 0.7819. The persistence benchmark achieved an MAE of 1.1913, indicating that it was a strong baseline for absolute error. Hence, the selected model is not claimed to outperform the persistence benchmark on all metrics



Relative Moisture Stress Index (RMSI)

A Relative Moisture Stress Index (RMSI) was formulated to interpret the predicted soil-moisture values relative to the distribution of target values in the training dataset. The predicted value is converted into a training-distribution percentile, and the relative stress value is calculated as:

$$\text{RMSI} = 100 - \text{Training-data percentile}$$

The RMSI categories used in the study are:

RMSI Range	Interpretation
0–10	Very Low Relative Stress
>10–20	Low Relative Stress
>20–40	Moderate Relative Stress
>40–60	High Relative Stress
>60	Very High Relative Stress

RMSI provides a relative interpretation of moisture stress and is not a direct measurement of crop water requirement or the exact quantity of irrigation water required.

IV. RESULTS

The machine-learning models were evaluated on 48 chronological test observations using MAE, RMSE, and R^2 . Linear Regression, Decision Tree, and Random Forest were compared across Feature Groups A, B, and C, as well as a persistence benchmark. Linear Regression with Group B achieved the best performance among the trained machine-learning models, with an RMSE of 1.7977 and an R^2 of 0.7819 and was selected for further SHAP and RMSI analysis.

Prediction Analysis

The selected model followed the observed next-day soil-moisture trend. However, some deviations were observed because the soil-moisture distribution in the test period was higher than that of the training period.

SHAP Analysis

SHAP was applied to explain the selected Linear Regression model. Current soil moisture was the most influential feature, followed by month sine, soil-moisture lag 3, soil-moisture lag 2, and soil-moisture lag 1. This indicates that current and historical soil-moisture conditions, as well as seasonal patterns, contributed to the predictions.

RMSI Analysis

RMSI was calculated using the training data distribution. Out of 48 test predictions, 33 were classified as Very Low Relative Stress, 14 as Low Relative Stress, and 1 as Moderate Relative Stress; none classified as High or Very High Relative Stress. RMSI provides a relative indication of moisture stress and does not represent an exact crop-water requirement.

Case Analysis

A low-error case on 2 September 2020 had an actual value of 33.9364 and predicted value of 33.9218, an error of 0.0146. A challenging case on 5 September 2020 had an actual value of 41.7935 and predicted value of 35.0590, an error of 6.7345, indicating that the available features could not fully capture this high soil-moisture condition.

V. DISCUSSION

The results indicate that historical soil-moisture, rainfall, and seasonal features can be used for next-day soil-moisture prediction. Linear Regression with Group B achieved the best performance among the trained machine-learning models based on RMSE and R^2 . However, the persistence benchmark achieved a lower MAE, indicating that the selected model was not superior to every metric. The addition of SHAP and RMSI provides explainability and relative moisture-stress information, supporting data-driven irrigation planning without specifying an exact amount of water.



VI. LIMITATIONS OF THE STUDY

The present research has several limitations that need to be considered. First of all, the data used in the study is dated, as it only includes information from 2020, with 240 observations in total. Secondly, the framework established only provides relative information on moisture stresses, which means that the amount of water required for irrigation cannot be determined. This is due to the lack of crop-specific, ontological, and root-zone variables in the model. Additionally, it is noteworthy that the authors used a small number of machine learning algorithms and environmental features. Finally, the difference between training and testing datasets might have impacted the accuracy of the results, as the framework might not perform well in case of extreme weather conditions. As such, the framework should be considered a proof-of-concept prototype that requires additional validation at a field level before any practical application.

VII. CONCLUSION

The study established an Explainable Machine Learning Framework for the evaluation of Soil Moisture Stress Assessment and Irrigation Decision Support using data set from Pune, Maharashtra. The Linear Regression model which used historical soil-moisture from the data set, rainfall, and seasonal features, had the highest prediction accuracy among the trained machine-learning models. The SHAP values were used to interpret the model and understand what factors influenced the predictions, while the RMSI values were used to determine the relative moisture-stress levels for irrigation decision support. Overall, the framework predicted, interpreted, and evaluated soil moisture-stress level for irrigation decision support without predicting an absolute amount of water.

REFERENCES

1. Alahmad, T., Neményi, M., & Nyéki, A. (2026). Explainable XGBoost-based models of root-zone soil moisture profiling using coupled Sentinel-2 and IoT data in loam and silt loam soil. *Discover Applied Sciences*, 8, 672. <https://doi.org/10.1007/s42452-026-08673-3>
2. Alahmad, T., Neményi, M., Széles, A., Ali, N., Hijazi, O., & Nyéki, A. (2025). Spatiotemporal prediction of soil moisture content at various depths in three soil types using machine learning algorithms. *Frontiers in Soil Science*, 5, 1612908. <https://doi.org/10.3389/fsoil.2025.1612908>
3. Bhoi, A., et al. (2021). IoT-IIRS: Internet of Things based intelligent-irrigation recommendation system using machine learning approach for efficient water usage. *PeerJ Computer Science*, 7, e578. <https://doi.org/10.7717/peerj-cs.578>
4. Challa, L. P., et al. (2024). Prediction of soil moisture using machine learning techniques: A case study of an IoT-based irrigation system in a naturally ventilated polyhouse. *Irrigation and Drainage*, 73(3), 1138–1150. <https://doi.org/10.1002/ird.2933>
5. Gu, Z., Zhu, T., Jiao, X., Xu, J., & Qi, Z. (2021). Neural network soil moisture model for irrigation scheduling. *Computers and Electronics in Agriculture*, 180, 105801. <https://doi.org/10.1016/j.compag.2020.105801>
6. Lamichhane, M., Mehan, S., & Mankin, K. R. (2025). Soil moisture prediction using remote sensing and machine learning algorithms: A review on progress, challenges, and opportunities. *Remote Sensing*, 17(14), 2397. <https://doi.org/10.3390/rs17142397>

