Drought Prediction using Landsat-8 Images and Remote Sensing

Mithulesh P.[1]*, Agilan B.[1], Sourav Kumar N.R.S. [1] and Dr Vani K.[1]

[1]*Department of Information Science and Technology, College of Engineering Guindy, Anna University, India.*

*mithunponraj03@gmail.com

Abstract: *Drought conditions often greatly impact the ecosystems and lands for agricultural purposes. This research endeavors to provide a systematic approach to predict the drought conditions of the land areas using the data from Famine Early Warning Systems Network Land Data Assimilation System and Soil Moisture Active Passive Mission which provide information about soil parameters. The former is used to obtain soil moisture values at 4 different ranges - 0 to 10, 10 to 40, 40 to 100, 100 to 200 centimeters. The latter is used to obtain the Soil Moisture Indices at Surface and Root Level. This study utilized Landsat-8 images of the various landsites in Tamil Nadu from 2015 to 2022. Once the images are collected, they are subjected to cloud masking and land region segmentation since the drought conditions of the land region are to be estimated. This is followed by extracting 11 bands from the satellite images. In addition to this, several thermal, vegetation and water indices are calculated to establish their relationship with soil moisture parameters. Finally, the above mentioned six soil moisture parameters are also collected for the same train sites to create a final dataset with six parameters as dependent variables to train Random Forest, AdaBoost and XGBoost model while the spectral bands and indices serve as independent features. The trained Random Forest model yielded an R-squared score of 0.78 outperforming the other models and was validated using K-fold cross validation. When a Landsat-8 image of the test site is provided for the trained model, it estimates the moisture values at four different levels and two Soil Moisture Indices at surface and root level. These Indices are then used to predict the drought levels at both surface and root zone while the 4 values at different depth ranges are used for analysis purposes.*

Keywords: Drought, Landsat-8, Random Forest, Soil moisture.

Introduction

Drought is a drastic condition that profoundly impacts society by disrupting socio-economic development through compromised food security and ecological imbalance in affected areas. Forecasting a drought at the earliest is very crucial for effective management and mitigation of their adverse effects. This study aims to assess drought conditions through the prediction of soil moisture indices at target location utilizing a robust approach which is an amalgamation of satellite imagery and advanced machine learning techniques. By providing high resolution data for analysis of temporal and spatial changes in soil moisture satellite imagery acts as a pivotal tool in this research. In addition to this satellites such as Soil

Moisture Active Passive (SMAP) yields us ground truth soil moisture data enabling the opportunity for precise drought assessments. Processing vast datasets, enhancement of the prediction accuracy, identify complex patterns and correlation between satellite derived variables and ground-truth observations are carried out through advanced machine learning techniques. Primarily this research aims at forecasting drought conditions accurately enabling proactive measures to mitigate their impact. In addition to this prediction of water content levels both at surface level and below the surface level through detailed soil moisture assessments. Heatmaps are also generated offering valuable insights into soil moisture dynamics across different depths, aiding in comprehensive drought monitoring and management strategies. This robust method provides real-time monitoring systems capable of updating drought forecasts based on the latest satellite observations of our target zone, leveraging the temporal resolution of satellite imagery to incorporate newer data for the newer assessments.

Literature Review

Climate change has boosted the unpredictability of droughts forecasted by historical meteorological data. Severe Drought Prediction model proposed by Haekyung Park et al (2019) tailored for short-term drought prediction, which integrates complementary data instead of conventional meteorological data. Four different categories of surface factors such as vegetation, topographic, water and thermal factors that can affect soil moisture were defined and 15 input variables from different categories were considered. A regression drought function was developed with Random Forest model providing a training performance of 0.91 R-squared score (R^2). Anurag Dash et al (2022) utilized Soil Moisture Index (SMI) taken from SMAP satellite and 12 other variables taken from Landsat-8 images to create a dataset. Aforementioned dataset was utilized in training of Random Forest model that yielded a training performance of 94.1%. Considering the performance given by Random Forest in both the researches implies Random Forest model is better suited for this task. However, the aforementioned approaches are very susceptible to meteorological anomalies since it is based on historical precipitation data. The choice of relying on Landsat-8 images for input resulted in mediocre data affected by cloud cover.

SMAP satellite mission was launched on January 31, 2015 providing global mapping of high-

resolution soil moisture and freeze-thaw state with temporal resolution of 2 to 3 days utilizing an L-band radiometer. In April 2015, SMAP initiated a program to gather both radar and radiometer data simultaneously. The primary objective was to generate three distinct soil moisture products: a radiometer-only product offering spatial resolution of 40km, a combined radar/radiometer product with a resolution of 10km, and a radar-only product providing details at a resolution of 3km. Results from Steven K Chan et al (2016) indicated that the Dual Channel Algorithm had the lowest bias, which was virtually zero relative to other soil retrieval algorithms. In the case of the V-pol Single Channel Algorithm, the bias was only 0.018 m^3/m^3 at core validation sites. Amy McNally et al (2017) employed the Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS), which is a customized version based on NASA's Land Information System (LIS). This system is regularly utilized to generate hydroclimate state estimates. The outputs from this initiative encompassed soil moisture percentiles and assessments of water availability. FLDAS facilitated the provision of both real-time operational estimates and high-quality research-oriented data through the operational and research arms of FEWS NET, thereby enhancing the monitoring capabilities of evolving environmental conditions.

The Random Forest algorithm, introduced by Leo Breiman in 2001, is a technique that utilizes a collection of decision trees, each trained on different subsets of the data and features. This approach, based on ensemble learning, has demonstrated notable improvements in classification accuracy by combining the predictions from multiple trees. Large number of trees are generated finding the most popular class by voting. With use of Strong Law of Large Numbers, it is proved that trees always converge avoiding the problem of overfitting. Injecting the right of kind of randomness makes them accurate classifiers and regressors making it suitable for our problem statement. Tree boosting stands as a highly effective and extensively utilized machine learning approach. Tianqi Chen et al (2016) introduced XGBoost, a scalable end-to-end tree boosting system that delivers cutting-edge solutions across various domains. It incorporates a novel sparsity-aware algorithm and employs a theoretically grounded weighted quantile sketch technique to enable approximate learning. By utilizing the aforementioned techniques XGBoost is able to solve real world problems with a smaller number of resources. The conversion of weak learners that perform slightly better than random guessing into strong learners with high accuracy through effective combination of weak hypotheses was a big challenge giving rise to a dynamic allocation problem. Additionally, the need of deriving bounds on the net loss incurred by the allocation

algorithm posed a big challenge too. To address these problems Yaov Freund et al (1997) proposed AdaBoost algorithm that provides systematic approach to enhance weak learners. AdaBoost employs a weighted distribution over the training samples allowing algorithm to focus more on the instances that are misclassified by previous weak learners. This dynamic adjustment of weights effectively allocates more resources to harder to classify thereby improving the learning process over time. AdaBoost combines multiple weak learners by summing their weighted predictions rather than using a majority vote. The algorithm offers theoretical guarantees for the final hypothesis's performance by ensuring that the aggregation of weak hypotheses effectively minimizes the overall error. AdaBoost's flexibility and effective handling of multi-class problems allow it to often outperform other boosting algorithms and traditional classifiers such as Support Vector Machine (SVM) in terms of accuracy.

Finding the right hyperparameters in machine learning is often resource intensive task. With the increasing trend in Deep Learning an optimal way of finding the better performing hyperparameter is the need of the hour. Takuya Akiba et al (2019) proposed Optuna, a next generate hyperparameter optimization framework. Optuna works by a deep learning philosophy named as define-by-run which provides user the privilege of not explicitly defining everything in advance about the optimization strategy. Optuna formulates the hyperparameters optimization as a task of minimizing/maximizing an objective function taking a set of hyperparameters as input and return the validation score. The objective function is gradually built through inter action with trial object with dynamic construction of search spaces of the trial object. Optuna establishes an efficient pruning algorithm by implementing a variant of Asynchronous Successive Halving (ASHA) in which early worker is allowed to asynchronously execute aggressive stopping based on provisional ranking of trials. The efficiency of Optuna framework becomes prominent when it became the key player in Preferred Networks' Faster Region-based Convolutional Neural Network (RCNN) models, High Performance Linpack (HPL), RocksDB and more.

Methodology

In this section, a step-by-step process on the Drought prediction is explained, as illustrated in Figure 1 for clarity and reference throughout the study.

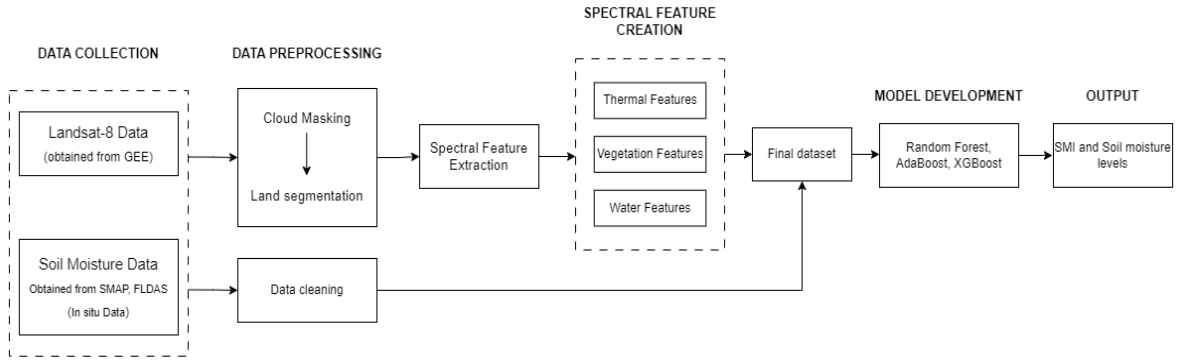

Figure 1: Overall architecture of the study

Data collection: The data source used in this study include SMAP- SMI at the root and surface levels from NASA/USDA-Enhanced SMAP retrievals. Another data involves the assessment of soil moistures through the FLDAS data, which provide measurements for 0-10 cm to 100-200 cm below the soil surface. These measurements have been taken from different coordinates within Tamil Nadu, spanning from the years 2020 to 2023. Together with these soil moisture data, Landsat-8 images were downloaded for the same locations and dates as the SMAP and FLDAS values. Google Earth Engine (GEE) is used for accessing data and setting up the data such that the satellite images are precisely spatially and temporally aligned for consistency and accuracy between different datasets.

Spectral Index calculation and final dataset preparation: Landsat-8 bands are used in deriving a number of key spectral indices for creating a comprehensive dataset with an objective for monitoring and prediction of drought conditions. Among the numerous spectral indices, one corresponds to the Normalized Difference Vegetation Index, NDVI. It reflects vegetation stress, and as such, it brings out areas of stressed vegetation, hence allowing for early detection and management of the impacts of droughts, and providing timeliness in intervention measures to avert devastating conditions on vegetation and farmlands.

Soil Adjusted Vegetation Index (SAVI) that improves upon NDVI with an input variable for soil brightness; that becomes the factor that increases the precision of the outcome in the effects of drought on vegetation health. Enhanced Vegetation Index (EVI) overcomes the influence of the atmosphere in the determination of NDVI, making it more exact in the estimation of the degree of drought and health of vegetation. Normalized Difference Moisture Index (NDMI) uses the Near Infra-Red (NIR) together with the Short-Wave Infrared bands to assess the moisture content of vegetation—a component which is key in both drought monitoring and prediction.

Modified Normalized Difference Water Index (MNDWI) developed through green and shortwave infrared (SWIR) bands, giving information relating to changes in water availability in the period during drought. Modified Soil Adjusted Vegetation Index (MSAVI)—improved sensitivity to the dynamism of vegetation health status—assists in locating drought patterns and impacts. Bands considered in this study include Red, Blue, Green, NIR, SWIR1, SWIR2, TIRS1, and TIRS2.

The information on variation in plant water content and vegetation stress is described by the NIR and the SWIR bands, while the indirect description of soil moisture dynamics is given by the TIRS bands and the thermal radiation. Also, NDWI use near real-time available information to estimate water availability, and, therefore, further shed light on the effect of drought on vegetation. All the bands were used, along with the derived spectral indices, as individual features in the final dataset. All formulas for all spectral indices are given in the Table 1.

Table 1. List of spectral indices along with their formulas

Spectral Indices	**Formula**
NDVI	$\frac{\text{NIR - Red}}{\text{NIR + Red}}$
SAVI	$\frac{\text{Red - Green}}{\text{(Red + Green +0.5) * 1.5}}$
EVI	$\frac{2.5 * (Red - Green)}{(Red + 6 * Green - 7.5 * Aerosol + 1)}$
NDMI	$\frac{NIR - SWIR1}{NIR + SWIR1}$
MSAVI	$\frac{2 * Red + 1 - \sqrt{(2 * Red + 1)^2 - 8 * (Red - Green)}}{2}$
MSI	$\frac{SWIR1}{NIR}$
NDWI	$\frac{NIR - SWIR1}{NIR + SWIR2}$
MNDWI	$\frac{Green - SWIR1}{Green + SWIR1}$

The culminating dataset incorporates indicators for soil moisture at both the surface level and root zone, as well as soil moisture measurements at various depth intervals: 0-10 cm, 10-40 cm, 40-100 cm, and 100-200 cm.

Feature Importance: Spectral indices and band reflectance values, extracted from remote sensing data, are essential for precise estimation of soil moisture content. In this paper, in order to estimate the efficacy, correlations of the spectral indices and bands with the soil moisture level and soil moisture index is calculated. The analysis outcomes demonstrated a strong affirmative relationship between indices sensitive to water content and the measured moisture levels in the soil. In this respect, it was the strongest with MNDWI, followed by NDWI: r = 0.354701 and r = 0.326438, respectively. This is supported by many previous studies indicating that these indices are sensitive to soil water content. Interestingly, the NDMI also indicated a positive correlation, r = 0.256321, which may suggest potential interactions between urban landscapes and soil moisture patterns. Vegetation-based metrics, notably NDVI, showed a positive link to soil water content levels. At the same time, SAVI and EVI were strongly negatively correlated, with r = -0.199466 and r = -0.232807 respectively. Specifically, analysis of the individual spectral bands indicates that the coastal aerosol band, B1 (0.43-0.45 μm), has the exhibited weak to moderate negative correlations with soil moisture with r = -0.045368. This is followed by the blue band, B2 (0.45-0.51 μm), which has a correlation coefficient of r = 0.176482. This probably means that short wavelengths of the visible spectrum are more sensitive to changes in soil moisture. The strongest negative correlation is given by the thermal infrared band, B11 (11.50-12.51 μm), which has an r = -0.552163. This disparity thus underlines the importance of wavelength selection in moisture assessment protocols, further underscoring, in general, complex interactions between soil moisture and electromagnetic radiation across different parts of the spectrum.

Model Development and Evaluation: Three Ensemble Learning regression models were used in training the dataset: Random Forest Regressor, AdaBoost Regressor and XGBoost Regressor. These models have strong mechanisms for dealing with large datasets because of the use of ensembles of decision trees, which are efficient ways of capturing complex patterns and greatly improving predictive accuracy. In this analysis, all hyperparameter tuning for the models was done using Optuna. Their performance is evaluated using standard metrics like R-squared and Root Mean Squared Error. The equations for RMSE and R^2 score are defined as follows.

The R^2 score is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Whereas RMSE is given by,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

The following metrics have provided quantitative measures of how good the models fit the data and predict.

Drought prediction: The measurement of the level of soil moisture at four depths helps to identify the availability of water and these measurements helps to assess the impact on agriculture productivity and health. The results predicted for the Soil Moisture Index and soil moisture levels over the area will be visualized, where the predicted values are attached to each pixel, hence, creating a drought map. This map categorizes the intensity of the drought at the surface and root levels and offers more detailed information about the drought impacts. This mapping will be very helpful in assessing and managing droughts with far more competencies at different depths and spatially. The SMI values are categorized as: 0 to 0.2, severe drought; 0.2 to 0.4, high drought risk; 0.4 to 0.8, low to very low drought risk; and from 0.8 to 1, minimal or no drought risk. This categorization helps in estimating the level and possibilities of drought over different regions.

Results and Discussion

The final dataset was approximately 13,234 data points, which were then randomly split into a training dataset and a testing dataset in the ratio 80:20. Random Forest, AdaBoost, and XGBoost were developed and calibrated using the prepared training dataset. The Hyper-Parameter tuning process using Optuna was then used to get the best performance. This process fine-tuned the models to improve their predictive accuracy on test data.

It tuned a comprehensive list of hyperparameters, launching 1000 trials to run through many different combinations in search of the best models. The process involved testing systematically all sets of different parameters in search of the best possible performance. Results for these trials for all of the models are summarized in Table 2, Table 3 and Table 4, including all tested hyperparameters, their respective performances, and the best parameters identified by Optuna. All of this tuning was a very integral part in the refinement of the model for better accuracy and efficiency.

Table 2. List of best Hyperparameter for Random Forest model obtained from Optuna

Hyperparameters	**Range for each Hyperparameter**	**Best Value obtained**
n_estimators	10 to 1000	469
max_depth	2 to 200	11
min_samples_split	2 to 50	5
min_samples_leaf	1 to 50	1
max_features	['auto', 'sqrt', 'log2', None]	sqrt
bootstrap	[True, False]	True

Table 3. List of best Hyperparameter for XGBoost model obtained from Optuna

Hyperparameters	**Range for each Hyperparameter**	**Best Value obtained**
lambda	1e-8 to 1.0	0.34888107520657696
alpha	1e-8 to 1.0	9.487580450505716e-06
max_depth	3 to 12	12
eta	1e-4 to 1.0	0.08840764127174541
gamma	1e-8 to 1.0	1.3437743943203425e-08
grow_policy	['depthwise', 'lossguide']	depthwise
subsample	0.5 to 1.0	0.6595273699333676
colsample_bytree	0.5 to 1.0	0.7941379176780416
min_child_weight	1e-8 to 1.0	2.1119948079914213e-07

Table 4. List of best Hyperparameter for AdaBoost model obtained from Optuna

Hyperparameters	Range for each Hyperparameter	Best Value obtained
max_depth	2 to 200	24
min_samples_split	2 to 50	25
min_samples_leaf	1 to 50	13
n_estimators	50 to 200	152
learning_rate	0.01 to 1.0	0.37808498449931166
loss	['linear', 'square', 'exponential']	exponential

Model performance was assessed using two metrics: R^2 score and RMSE. To make this more robust in terms of evaluation, the performance assessment was made using K-fold cross-validation. The data is divided into 5 subsets, and models are iteratively trained and tested on these subsets. This will ensure robust model evaluation, avoiding overfitting risks, and gives the full view of the predictive capabilities of each model considered. Table 5 present, R^2 score and RMSE values for the Random Forest, AdaBoost and XGBoost model, respectively. It provides a detailed assessment of their accuracy and consistency across various data splits.

Table 5. Performance of each Models

Model	R^2 score	RMSE
Random Forest	0.78295	0.0384
XGBoost	0.7619	0.0397
AdaBoost	0.7343	0.0408

The Random Forest model has outperformed the AdaBoost and XGBoost model. Therefore, Random Forest has been selected for making predictions. The places selected for study are Puzhal and Theri Kaadu. The test site chosen near Puzhal Lake - land beside one of the largest lakes in Tamil Nadu—was opted for, while Theri Kaadu, situated in Southeast Tamil Nadu, represents a red soil desert condition zone. These sites are chosen to assess the accuracy of the model in two contrasting different environments. Landsat-8 images from these two locations were downloaded from GEE. The test images are subjected to cloud masking and land segmentation. Cloud masking is done with the help of Quality Assurance Band (QA_Band). Land segmentation is carried out with the view of masking out water

bodies from the image. Then the resultant image is used as input to the Random Forest model, and outputs were derived. Figure 2 represents the output for Puzhal, while Figure 3 represents the output for Theri Kaadu.

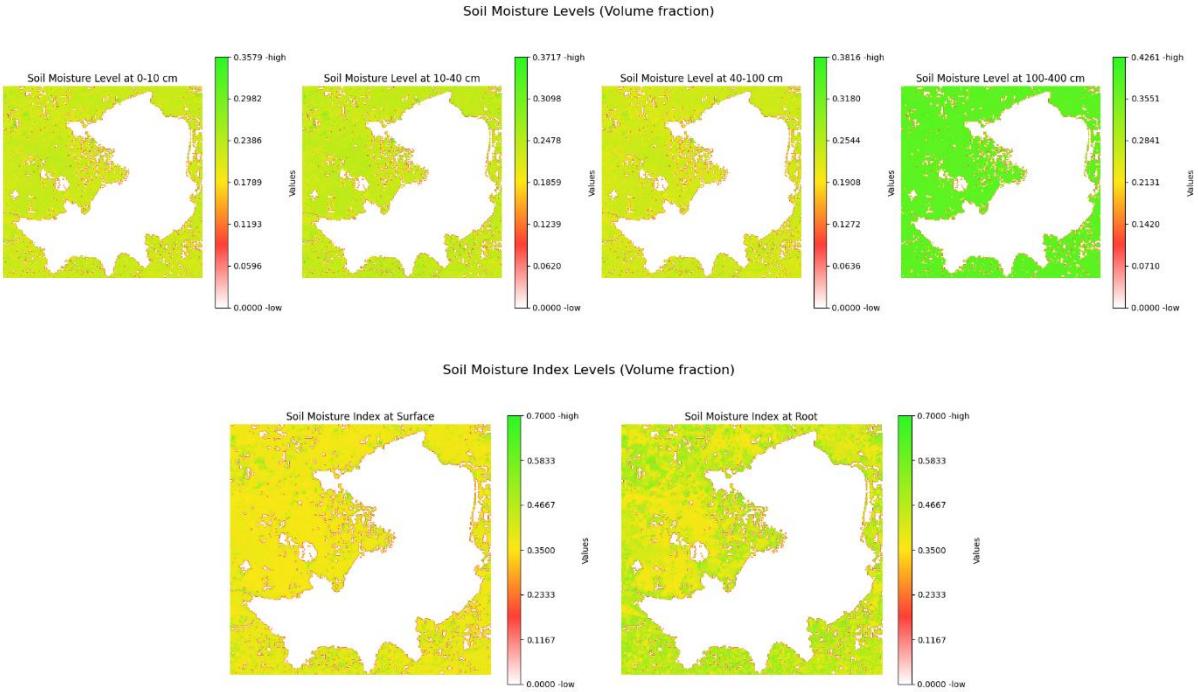

Figure 2. Soil moisture levels and SMI values of Puzhal.

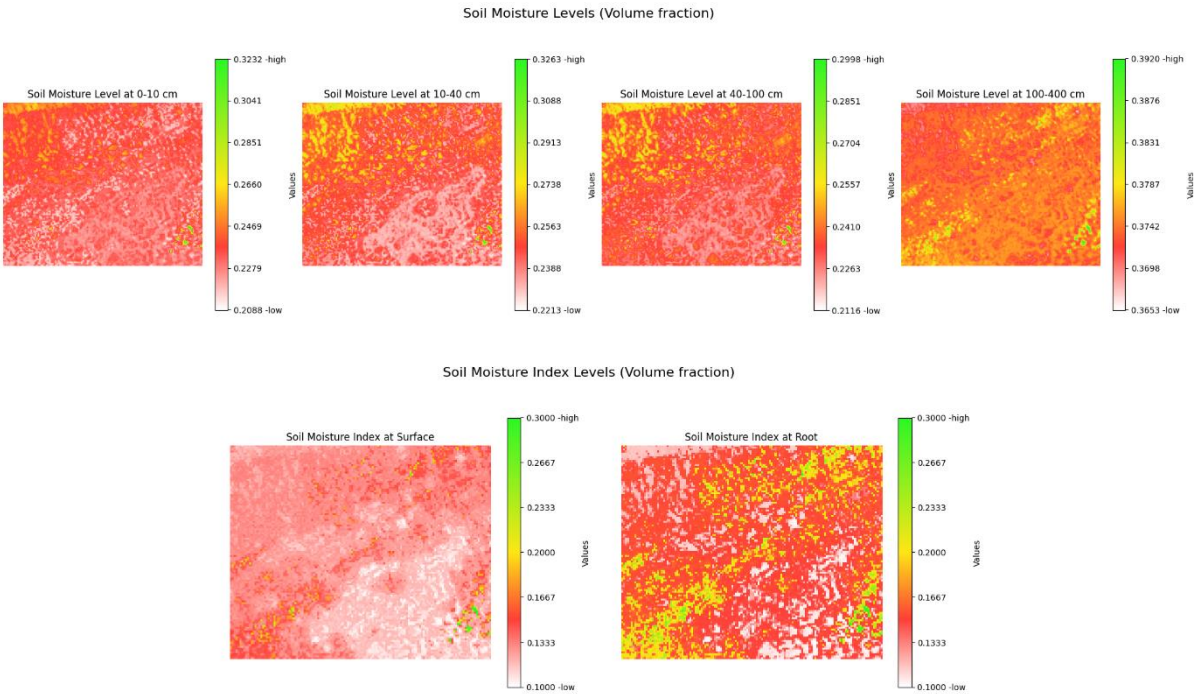

Figure 3. Soil moisture levels and SMI values of Theri Kaadu.

The results show the soil moisture levels and soil moisture index for two test sites. Green means high presence, red means low presence, and white indicates water or no data. From the Figure 2, it is obvious that the land adjacent to Puzhal lake contains high soil moisture levels, where the mean soil moisture indices are 0.42 at the surface zone and 0.57 at the root zone. Thus, this site is described as having a "low to very low risk of drought". Soil moisture levels at each depth also indicate higher water availability throughout the profile.

In sharp contrast, Figure 3 shows that, at Theri Kaadu, the soil moisture levels are very low. The mean SMI values here range from 0.131 at the surface zone to 0.29 at the root zone. At the surface level, this places it at "severe drought," while at root level, it is "high drought risk". Soil moisture at individual depths indicates reduced water availability along the profile. Across different sites, this analysis brings into relief the variability of the drought conditions that were modelled, showing how the water availability at Theri Kaadu was different from that at Puzhal Lake, and how their respective drought severities differed.

Conclusion and Recommendation:

The result of this research confirms the effectiveness of integrating satellite-derived data with machine learning techniques in drought prediction and monitoring. In this study, SMAP and FLDAS soil moisture data were used together with Landsat-8 spectral bands and indices to develop a robust Random Forest model with an accuracy of 78% able to suitably assess drought conditions. The model was tested against different scenarios, and its efficiency has been proved by rigorous cross-validation. The case studies of Puzhal and Theri Kaadu represent an important part of model capacity, providing an efficient way of knowing the varying risks of drought. This kind of granular assessment on a surface and root-zone level gives very useful information on agricultural planning and water resource management. This approach is very useful because scalable and adaptive solutions for drought monitoring are essential, particularly in regions with limited ground-based data. The methodology combines satellite data, which is easily available, with machine learning algorithms to come out with a promising tool that will help in preparedness and mitigation efforts concerning droughts. Future research may consider geographical scale enlargement and the addition of more environment variables. Making the model more responsive to rapidly changing conditions can be achieved by integrating real-time data streams. Deep learning techniques can be explored for applicability, and test the performance of the model in different climatic zones to set up a more accurate and of wider applicability of the drought prediction. Hence this

research study describes the potential of advanced data analytics and artificial intelligence in remote sensing, enabling any entity to enhance its predictive capabilities and conduct accurate environmental monitoring. These are, therefore, techniques that demonstrate how technology might help shift insight and management of our ecosystems by offering a method for the right analysis and prediction of data.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019*). Optuna: A next-generation hyperparameter optimization framework*. arXiv. https://doi.org/10.48550/arXiv.1907.10902

Breiman, L. (2001). *Random forests. Machine Learning, 45(1),* 5-32. https://doi.org/10.1023/A:1010933404324

Chan, S. K., Bindlish, R., O'Neill, P. E., Njoku, E., Jackson, T., Colliander, A., Chen, F., Burgin, M., Dunbar, S., Piepmeier, J., Yueh, S., Entekhabi, D., Cosh, M. H., Caldwell, T., Walker, J., Wu, X., Berg, A., Rowlandson, T., Pacheco, A., ... Kerr, Y. (2016). *Assessment of the SMAP passive soil moisture product. IEEE Transactions on Geoscience and Remote Sensing, 54(8), 4994-5007.* https://doi.org/10.1109/TGRS.2016.2561938

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).* ACM. https://doi.org/10.48550/arXiv.1603.02754

Dash, A., Jetley, S., Rege, A., Chopra, S., & Sawant, R. (2022). *Drought prediction and water quality estimation using satellite images and machine learning. In 2022 7th International Conference on Communication and Electronics Systems (ICCES) (pp. 1110-1116).* IEEE. https://doi.org/10.1109/ICCES54183.2022.9835727

Freund, Y., & Schapire, R. E. (1997). *A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1),* 119-139. https://doi.org/10.1006/jcss.1997.1504

McNally, A., Arsenault, K., Kumar, S., Shukla, S., Peterson, P., Wang, S., Funk, C., Peters-Lidard, C. D., & Verdin, J. P. (2017). *A land data assimilation system for sub-Saharan Africa food and water security applications. Scientific Data, 4,* 170012. https://doi.org/10.1038/sdata.2017.12

Park, H., Kim, K., & Lee, D. K. (2019). *Prediction of severe drought area based on random forest: Using satellite image and topography data. Water, 11(4),* 705. https://doi.org/10.3390/w11040705

Zhang, H.-w., & Chen, H.-l. (2015). *The application of modified normalized difference water index (MNDWI) by leaf area index in the retrieval of regional drought monitoring. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XL-7/W3,* 141-147. https://doi.org/10.5194/isprsarchives-XL-7-W3-141-2015