International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Moisture Stress Detection in Soybean Crops Using Sentinel-2 Time-Series NDVI and Machine Learning Techniques

Rahul B. Mannade

Submitted:03/11/2021 **Revised**: 14/12/2021 **Accepted:** 25/12/2021

Abstract: The measurement of moisture stress in crops is necessary in enhancing agricultural productivity as well as making water resource management sustainable especially in rainfed agricultural systems. Conventional field-based approaches tend to be restricted in spatial and temporal terms, and remote sensing could be considered as an option to monitor crops on a large scale. This paper introduces a machine learning-based model of moisture stress detection in soybean plants with the help of time-series satellite data. Sentinel-2 multispectral imagery was used in the growing season. Normalized Difference Vegetation Index (NDVI) and Normalized Difference Water Index (NDWI) were calculated using monthly composite images to calculate vegetation and moisture indices respectively. The dataset was in pixel format with moisture stress labels determined based on NDWI values that were determined by a dynamic threshold method. To prevent leaking of data and provide model reliability, NDVI time-series characteristics were taken as input variables, whereas NDWI was applied only to label generation. A Random Forest model was used to estimate the connection between the vegetation dynamics and the moisture stress situation. The overall accuracy of the model was 93.33, which means that the model has a high predictive power. The analysis of the importance of features showed that the values of NDVI during the later growth stages and especially in September and October were the most significant to detect the stress, and crop monitoring over time is important. The findings indicate that vegetation indices are useful to monitor the patterns of moisture stress without necessarily using water-related inputs. This research establishes the possibility of combining remote sensing and machine-learning methods in an efficient and scalable crop stress monitoring. The suggested solution offers an economical solution in making agricultural decisions and can be generalized to other crops and area to apply precision farming.

Keywords: *NDVI, NDWI, Time Series NDVI, ML*

1. Introduction

In areas where crop productivity is very sensitive to climatic factors, and water, agriculture is critical to ensure food security. There are other environmental factors such as soil moisture, which has a profound effect on crop growth, development and yield [1]. In rainfed agricultural systems like most of India, the moisture stress cannot be monitored and thus to make timely decisions and to manage crops sustainably. Conventional water stress measurements of crops are traditionally labor intensive, time consuming, and have a small spatial coverage and are therefore less effective in monitoring large areas [2].

The current development in remote sensing systems has given potent means of monitoring agriculture in large scale. The frequent revisit time and high-resolution multispectral images of satellites like Sentinel-2 provide the opportunity to constantly monitor the situation with crops [3]. The use of vegetation indices based on satellite data, especially the Normalized Difference Vegetation Index (NDVI), has gained wide use in measuring crop health, vigor and biomass. Moreover, water indices like the Normalized Difference Water Index (NDWI) can give information on the state of moisture in the crops, and are thus useful in the study of stress detection [4].

Department of Information Technology, Government College of Engineering, Aurangabad, Maharashtra, 431005, India
mannade.rahul@gmail.com

Machine learning methods have also added to the ability of remote sensing data analysis in that they allow automated and precise crop condition classification. Some algorithms like the Random Forest have become popular because of their strength, capability to deal with non-linear relationships as well as their performance with high in dimensional data [5]. By combining time-series satellite with machine learning solutions, one will be able to detect temporal trends in crop development and stress situations with better precision and scalability, resulting in more reliable and scalable agricultural monitoring systems [6].

Soybean is a type of oil seed crop that is widely grown in India especially in the Kharif season. It is very sensitive to water and its growth is affected by the water availability in certain critical periods of growth like flowering and pod development [7]. During these stages, moisture stress may result in a large decrease in yield and quality. Early observation and monitoring of water stress on soybean fields is thus important in maximising irrigation and enhancing crop yield. Such monitoring can be effectively done through remote sensing based techniques which are cost effective [8].

The NDVI was taken as the main input feature, and the information derived by means of NDWI was applied to obtain stress labels, avoiding the data leakage. The dependence between vegetation dynamics and moisture stress conditions was modeled with the help of a Random Forest classifier. The findings show that vegetation indices are useful in indirectly detecting moisture stress and that late-season crop indicators are important to enhance the classification accuracy.

2. Literature Survey

The latest developments in remote sensing and machine learning have made a lot of progress in the ability to track the health status of crops and identify the presence of moisture stress with the help of satellite imagery. A number of papers released have delved into the incorporation of vegetation indices and data-driven techniques to be utilized in agriculture.

One of the authors conducted a study to examine how multi-temporal Sentinel-2 data can be used to monitor crop conditions. The vegetation indices used by the authors included NDVI and enhanced vegetation index (EVI), to examine the dynamics of crop growth at various phenological stages. Their findings indicated that time-series NDVI is very effective in identifying temporal changes in crop health and could be utilized as a predictable marker of detection of stress conditions. The research noted that, when used correctly, time-based data is crucial in aiding proper agricultural evaluation compared to one-day images [9].

Author focused is another contribution, which focuses on crop water stress detection based on spectral indices based on Sentinel-2 imagery. The research also mentioned the importance of the water-related indices like the NDWI in detecting changes in moisture in crops. Using spectral characteristics and machine learning methods, the authors obtained higher classification rates between stressed and non-stressed crop states. Their results confirm the combination of spectral indices with data-driven models to enable effective stress monitoring [10].

Additionally, the use of machine learning methods, specifically Random Forest, to classify crops and evaluate their condition using multispectral satellite images, was presented by another author. The study published in MDPI Sensors and revealed that Random Forest is better than the traditional classification methods because it can capture complex relationships and noisy data. Another important point that the authors made concerns the importance of feature importance analysis, which can be used to determine the most influential variables that are going to be used to determine the state of crops, which, in this case, are vegetation indices [11].

All in all, these findings suggest that the combination of Sentinel-2 time-series data, water and vegetation indices, and machine learning algorithms is a powerful model to monitor crops and detect moisture stress. Based on these findings, the current research is on soybean crops and it utilizes the NDVI-based temporal characteristics alongside machine learning algorithms to detect moisture stress without the problem of data leakage.

3. Methodology

For methodological research we consider sillod area of Chhatrapati Sambhajinagar, Maharashtra, India. Figure 1 shows proposed methodology for our study.

Figure 1: Proposed Methodology

3.1. Data Collection

Sentinel-2 offers a high spatial resolution (10 m), and frequent revisit, which is why it is very appropriate in agricultural research. To ensure a precise analysis, the area of interest (AOI) was defined as farm boundary shapefiles that represented the fields of soybean and all satellite data were restricted to the AOI.

3.2. Preprocessing

Preprocessing step included sifting out the satellite imagery using time/atmospheric factors. Pictures in the given time frame were observed and those pictures that had too much cloud cover were discarded to reduce noise and distortion of data. The rest of the images were trimmed to the AOI and monthly median composites were formed in each of the months between July and October. Median compositing was used to minimize the impact of residual clouds and atmospheric variations to achieve cleaner and more confident datasets to be further analyzed [12].

3.3. Vegetation and Moisture Indices Extraction

Spectral indices were derived to reflect vegetation health and moisture conditions after preprocessing. Normalized Difference Vegetation Index (NDVI) was estimated to determine crop vigor and biomass whereas the Normalized Difference Water Index (NDWI) was estimated to determine the content of water and the availability of moisture in the vegetation [13]. Each of the monthly composites produced these indices, so as to produce a time-series dataset of NDVI and NDWI values (for July, August, September, and October). This time-based data is essential in comprehending crop development patterns and stress patterns [14].

3.4. Dataset Preparation

The stacked index images were sampled pixel-by-pixel to create a dataset to be analyzed using machine learning. The pixels inside the AOI were considered as independent observations and the respective values of NDVI and NDWI of all the months were taken. This method allowed creating a big enough dataset, which was later exported as a CSV file to be further processed within a machine learning setting. The pixel-level data allows greater variability and strength of the data set than aggregated field-level statistics.

3.5. Label Generation (Moisture Stress)

The NDWI values were used to create moisture stress labels to define the target variable. The average NDWI of all months was calculated at each pixel and a dynamic threshold calculated using the median value was used to categorize the data into two groups; stress and no stress. Pixels whose values of NDWI were less than the threshold were classified as moisture-stressed whereas those whose values were greater were classified as non-stressed. In order to prevent the leakage of data, and offer a fair analysis of the model, NDWI was only utilized in the generation of labels, but it was not added as an input feature when training the model.

3.6. Feature Selection

The model used NDVI time-series variables as input features to select the features. This choice made sure that the model acquired knowledge of the interaction between the vegetation dynamics and moisture stress contingent upon the indirect manner, which is not based on direct indicators of moisture. The features that were selected were the NDVI values in July, August, September, and October, which are at various growth stages of the soybean crop.

3.7. Model Development: Random Forest

In the classification task, a random forest algorithm was used because it is robust, capable of managing non-linear relationships and it performs well in remote sensing tasks [15]. The data was classified into training and test sets where 70 percent of the data was utilized to train the model and the rest 30 percent was used to verify the model. The NDVI features were used to train the model to categorize pixels as either stress or non-stress.

3.8. Model Evaluation

Standard classification measures such as accuracy, confusion matrix, precision, recall and F1-score were used to evaluate the performance of the trained model. These measures were a complete evaluation of the predictive power of the model [16]. Also, the importance of its features was analyzed to determine the role that each NDVI variable played in predicting moisture stress. The analysis aided in comprehending the growth stages that had the greatest influence in the identification of stress conditions.

3.9. Result Analysis

Lastly, the findings were discussed to explain the spatial and temporal variations of moisture stress of soybean crops. The contribution of late-season vegetation indices was of specific interest as such indices tend to be more reflective of the cumulative effects of stress. The results emphasize the usefulness of combining remote sensing data with machine learning methods to achieve effective and scalable crop stress monitoring.

4. Result and Discussion

The standard classification metrics such as accuracy, confusion matrix, precision, recall, and F1-score were used to determine the performance of the proposed Random Forest model in moisture stress detection in soybean crops. The time-series features of NDVI were used to train the model and test on unknown test data, providing a good estimate of predictive power. The model had a total accuracy of 93.33 which implies that there is a high degree of agreement between the predicted and actual moisture stress conditions. This finding confirms that vegetation indices based on the Sentinel-2 imagery can be very useful in the detection of crop stress patterns even without considering the water-related indices in the model input.

Actual \ Predicted	No Stress (0)	Stress (1)
No Stress (0)	8	0
Stress (1)	1	6

Table 1: Confusion Matrix

The confusion matrix shows in table 1, that the model perfectly classified all the non-stressed samples (8 out of 8) which is the perfection of the class of No Stress. Nevertheless, a single stressed sample was incorrectly categorized as non-stressed, which means that it is slightly limited in identifying some stress states. Nevertheless, the general performance of the classification is good and there is only one misclassification among 15 samples.

Class	Precision	Recall	F1-Score	Support
No Stress (0)	0.89	1	0.94	8
Stress (1)	1	0.86	0.92	7
Overall Accuracy	—	—	**0.93**	**15**
Macro Avg	0.94	0.93	0.93	15
Weighted Avg	0.94	0.93	0.93	15

Table 2: Classification Metrics

According to the classification report represented in table 2, the model attained a precision of 1.00 on the stress class, which implies that all of the predicted stresses were accurate, and none of them were false positives. The stress class recall is however a little lower (0.86) indicating that there was probably one real stress case that was not observed. The F1-score of each of the two classes is greater than 0.90 which attests to a balanced and dependable model performance.

Feature	Importance
NDVI_Oct	0.426
NDVI_Sep	0.324
NDVI_July	0.169
NDVI_Aug	0.081

Table 3: Feature Importance Analysis

The analysis of the importance of the features indicates as shown in table 3, the relative role played by each NDVI variable in the prediction of moisture stress. Based on the results, it is evident that NDVI_Oct (42.6%) and NDVI_Sep (32.4) are the most significant features, which bring about almost three-quarters of the total significance. Conversely, earlier indices like NDVIJuly and NDVIAug have isotropic contributions.

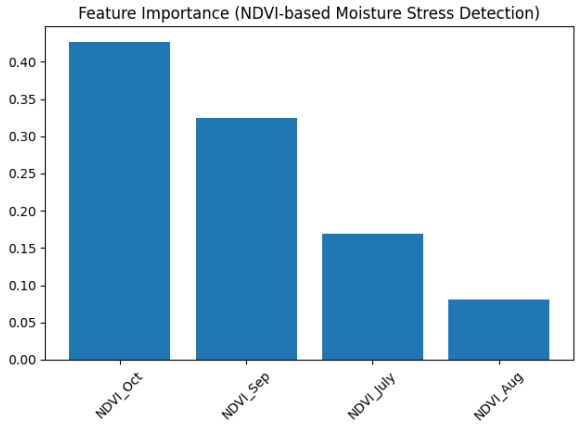

Figure 2: Feature Importance Graph

The feature importance graph shows in figure 2, the relative role of each NDVI time-series variable in predicting the moisture stress conditions. As can be seen, NDVI_Oct and NDVI_Sep prevail in the model and have the highest importance scores, implying that the conditions of the vegetation at the late stages of growth is the most significant predictor of moisture stress in soybean crops. Conversely, indices like NDVI_July and NDVI_Aug were found to be less significant in early years and this may indicate that the impact of stress factors is less intense during the early growth stages. In general, this graph demonstrates that the dynamics of vegetation as a time-dependent process is important, especially in its later phases to be used as a reliable stressor.

Discussion

The predominance of late-season NDVI components indicates that the impact of moisture stress is enhanced in later stages of the soybean crops growth. This finding is consistent with agronomic knowledge, where water stress has been found to severely affect crop physiology and yield during such critical phases of growth like flowering and pod development. This means that vegetation signals recorded in September and October give better indication of the stress conditions. The classification accuracy obtained in this study is high, which shows that the NDVI time-series data can be effectively used to detect indirect moisture stress. Using NDWI only as a label generation and omitting it as an input feature, the study manages to prevent data leakage as the model does not learn to correlate but to learn meaningful relationships. This method will improve the validity and external applicability of the findings.

The small difference in recall of the stress class suggests that not all the stress conditions could be effectively captured in vegetation indices especially the mild or early stress conditions. This implies that the detection accuracy can be further enhanced by incorporating other sources of data, including thermal imagery or soil moisture data. In general, the findings support the claim that machine learning models with the use of Sentinel-2-based vegetation indices offer a scalable and powerful solution to crop moisture stress detection. The methodology is specially applicable to large-scale agricultural applications, where quick and cost-efficient determination of crop conditions is needed.

5. Conclusion

This paper introduced a machine learning-based model to identify moisture stress in soybean plants using Sentinel-2 time-series data. The proposed methodology was able to prevent information leakage in the form of data leakage and achieve a high predictive accuracy by utilizing information of vegetation derived using NDVI and employing NDWI solely to produce labels. The overall accuracy of the Random Forest model was 93.33, which proves that satellite-derived vegetation indices are effective in detecting moisture stress conditions. Its findings indicate the significance of time analysis, and it can be concluded that late-season NDVI (September and October) are the most significant predictors in stress detection. This observation can be attributed to agronomic knowledge that moisture stress is enhanced at important growth phases including flowering and pod development. The paper validates the claim that NDVI can be used as a valid proxy of moisture stress when paired with the use of suitable machine learning methods.

Moreover, the suggested framework provides a scalable and economical method of monitoring large areas of agricultural activities, especially in areas

with limited ground-based observations. Remote sensing and machine learning can be used to assess the condition of crops at the appropriate time, thereby assisting in decision-making about the management of irrigation and how to increase yield. Nonetheless, the study is restricted with a single-season data and the lack of other environmental factors like soil moisture or temperature. Future studies can be directed at using multi-year data, adding more spectral and climatic variables, and using more advanced deep learning architectures to further improve prediction accuracy and the generalization of results. Altogether, the results prove the promise of the Sentinel-2 imagery and machine learning methods as an effective tool to monitor moisture stress in soybean farming and develop precision farming practices.

6. References

[1] Pandey, G. (2018). Challenges and future prospects of agri-nanotechnology for sustainable agriculture in India. Environmental Technology & Innovation, 11, 299-307.

[2] Wijewardana, C., Reddy, K. R., & Bellaloui, N. (2019). Soybean seed physiology, quality, and chemical composition under soil moisture stress. Food chemistry, 278, 92-100.

[3] Kobayashi, N., Tani, H., Wang, X., & Sonobe, R. (2020). Crop classification using spectral indices derived from Sentinel-2A imagery. Journal of Information and Telecommunication, 4(1), 67-90.

[4] Zheng, Y., Tang, L., & Wang, H. (2021). An improved approach for monitoring urban built-up areas by combining NPP-VIIRS nighttime light, NDVI, NDWI, and NDBI. Journal of cleaner production, 328, 129488.

[5] Gao, J., Nuyttens, D., Lootens, P., He, Y., & Pieters, J. G. (2018). Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery. Biosystems engineering, 170, 39-50.

[6] Elavarasan, D., & Vincent, P. D. R. (2021). A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. Journal of Ambient Intelligence and Humanized Computing, 12(11), 10009-10022.

[7] Jha, P. K., Kumar, S. N., & Ines, A. V. (2018). Responses of soybean to water stress and supplemental irrigation in upper Indo-Gangetic plain: Field experiment and modeling approach. Field crops research, 219, 76-86.

[8] Liu, Y., Kumar, M., Katul, G. G., Feng, X., & Konings, A. G. (2020). Plant hydraulics accentuates the effect of atmospheric moisture stress on transpiration. Nature Climate Change, 10(7), 691-695.

[9] Chen, J., et al. (2019). Monitoring Crop Condition Using Multi-Temporal Sentinel-2 Data. Remote Sensing, MDPI.

[10] Zhang, X., et al. (2020). Crop Water Stress Detection Using Spectral Indices and Machine Learning. Remote Sensing, MDPI.

[11] Xu, Y., et al. (2020). Crop Classification Using Random Forest and Multispectral Satellite Data. Sensors, MDPI.

[12] Mandal, B., Dolui, G., & Satpathy, S. (2018). Land suitability assessment for potential surface irrigation of river catchment for irrigation development in Kansai watershed, Purulia, West Bengal, India. Sustainable Water Resources Management, 4(4), 699-714.

[13] Ashok, A., Rani, H. P., & Jayakumar, K. V. (2021). Monitoring of dynamic wetland changes using NDVI and NDWI based landsat imagery. Remote Sensing Applications: Society and Environment, 23, 100547.

[14] Jothimani, M., Gunalan, J., Duraisamy, R., & Abebe, A. (2021, September). Study the Relationship Between LULC, LST, NDVI, NDWI and NDBI in Greater Arba Minch Area, Rift Valley, Ethiopia. In 3rd International conference on integrated intelligent computing communication & security (ICIIC 2021) (pp. 183-193). Atlantis Press.

[15] Chafik, H., Berrada, M., Legdou, A., Amine, A., & Lahssini, S. (2020, May). Exploitation of spectral indices NDVI, NDWI & SAVI in random forest classifier model for mapping weak rosemary cover: Application on Gourrama region, Morocco. In 2020 IEEE international conference of Moroccan geomatics (Morgeo) (pp. 1-6). IEEE.

[16] Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems* (pp. 79-91).