International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 18 (2017) pp. 7169-7175

A Hybrid of Plant Leaf Disease and Soil Moisture Prediction in Agriculture Using Data Mining Techniques

D. Sabareeswaran
Research Scholar, Department of Computer Science,
Karpagam Academy of Higher Education (KAHE), Karpagam University, Coimbatore, Tamilnadu, India.

Orcid Id: 0000-0002-3121-0946

Dr. R. Guna Sundari
Associate Professor, Department of Information Technology,
Karpagam Academy of Higher Education (KAHE), Karpagam University, Coimbatore, Tamilnadu, India.

Orcid Id: 0000-0003-4157-285X

Abstract

This paper focuses on developing novel technologies for monitoring the plant health, evaluating the water content level for different plants by analyzing the plant and soil images respectively. Initially, plants and soil images are captured through digital camera with a required resolution. The inner-distance shape context based descriptor and geometrical descriptors are used for extracting the shape and geometric features from the plant images. In addition, texture and color features are extracted from the soil images. By using the contour features of the plant images, the plant type is identified through the botanical plant species dictionary. The leaf diseases are predicted by Transductive Support Vector Machine classification. The causes for the specific plant disease are identified based on the Latent Dirichlet Allocation and Artificial Neural Network classification technique through the features of soil images and the diseased plant images. The obtained results are broadcasted to the cultivators through mobile phones by means of text messages on a daily and seasonal basis with possible suggestions of preventive measures.

Keywords: Crop production, Shape descriptors, Geometrical descriptors, Plant disease, Latent Dirichlet Allocation, Transductive Support Vector Machine, Artificial Neural Network.

INTRODUCTION

Agriculture is the backbone of India as well as the entire world. The evolution of agricultural technologies is engaged millions of years back and its advancement has been encouraged and defined through different environments, civilizations and technologies. In previous years, agriculture was grouped by improved yield, the replacement of man-made fertilizers and insect repellents for environmental pollution and land allowance. In modern lifetime, the environmental causes are removed hence macrobiotic and sustainable agriculture activities are developed.

The largest agricultural productivity with the highest quality is achieved through recent prediction techniques which provide the appropriate intensification environment under reproduction conveniences protection circumstances [1]. The modern growth of agricultural activities is monitored by the data mining techniques and enhanced information and communication processes [2]. Normally, the data mining is described as the process of realizing samples from large databases. The main objective of the data mining techniques is to mine the most significant features from database and convert the extracted features into a reasonable formation for additional utilize.

The data mining techniques [3] in agriculture are varied based on the agricultural applications. Some of the data mining techniques are provided for forecasting the environmental conditions such as weather conditions, environmental pollutions and etc. Some of the data mining techniques are utilized for recognizing purposes such as soil characteristics, weed detection and also used for monitoring water cores. However, the entire crop yield management using novel technology is the most significant issue in agriculture.

Hence in this paper, an effective data mining technique is proposed for predicting the required water content level for plants and different types of diseases. The major role of this paper is to identify the types of plant diseases and its causes for different plants. The plant diseases are predicted by the TSVM classification method. Also, the causes of the plant diseases are predicted based on the soil image features using ANN classification which uses the probability values between features of soil images and diseased plant images. Furthermore, the predicted parameters are forwarded to the farmers through messages or emails consistently on a daily and seasonal basis along with considerable precautionary recommendations for further improvement.

The remainder of the article is organized as follows: Section 2 describes about the different data mining techniques applied in agricultural fields. Section 3 explains about the proposed crop yield monitoring system using novel data mining techniques. Section 4 illustrates the performance evaluation of the proposed techniques. Section 5 concludes the research work and also provides the future work for further improvement.

LITERATURE SURVEY

A novel optimized spectral index [4] was developed to identify and forecast the winter wheat diseases. The various pests in winter wheat like powdery mildew, yellow rust and aphids were studied. The novel optimized spectral indices were determined by the weighted mixed single band and normalized wavelength variation of two bands. Initially, the majority and minority significant wavelengths were extracted from leaf spectral information for several diseases by using RELIEF-F algorithm. The reflectance of single band was extracted from the significant wavelengths and normalized wavelength variation from all the probable mixtures of majority and minority significant wavelengths were utilized for obtaining the optimized spectral indices. Moreover, the binary classification was performed to identify healthy plants and diseased plants. However, the sensitivity of disease detection was less also the detection methods were required to detect disease in earlier stages.

An operational agricultural drought monitoring [5] was developed by evaluating the utility of remotely sensed soil moisture retrievals. An ensemble Kalman filter data assimilation technique was developed for integrating surface soil moisture retrievals from NASA advanced microwave scanning radiometer (AMSR) into the United States Department of Agriculture (USDA) modified Palmer soil moisture model. The surface soil moisture dynamics evaluated by AMSR-E was utilized for updating the root-zone soil moisture estimation indirectly through the vertical soil moisture coupling of 2-layer soil moisture model. However, the most improved global soil moisture observations were needed for further improvement in accuracy.

The plant leaf disease symptom prediction algorithm [6] was introduced by using semi-automatic segmentation technique. The algorithm was performed based on the grayscale histograms to differentiate signs and symptoms of leaf diseases from asymptotic tissues in leaves. The histograms of H and a* were acquired from HSV and L*a*b* color space respectively by using this algorithm. The different issues of in-depth analysis were studied in this paper. However, the issue with this segmentation algorithm was that the accuracy was reduced since some sources of errors were not eliminated.

An automatic detection technique [7] based on computer vision was proposed for detecting crop diseases. The combination of marker controlled watershed and super-pixel based segmentation was performed for extracting the features. The feature selection process was achieved according to the textural, Gabor, gradient and biologically inspired features. After, the Support Vector Machine (SVM) was introduced for feature classification process and compared with the ANN classification method. The complexity of image processing functions was reduced by this technique. But, this technique has less robustness and collects only a certain amount of information from various symptoms.

The soil moisture prediction technique [8] was proposed by using improved BP algorithm. The time series of soil moisture information obtained from the wireless sensor networks were predicted based on the proposed BP neural network and particle swarm optimization algorithm. Initially, the time series of BP parameters were derived. Particle swarm optimization algorithm was introduced to improve the weight and threshold value of BP. Thus, the time series of soil moistures were predicted by the BP-PSO algorithm. However, the BP-PSO algorithm has high mean square error.

The agricultural monitoring system [9] was improved by using ant colony optimization with centre data aggregation algorithm. Generally, all sensors were joined together into weather stations and forecasting nodes were deployed in the data aggregation process. Here, an energy efficient centre data aggregation technique was proposed, where ant colony optimization algorithm was concerned with the manufacture of gradient field level. The crop production was monitored by developing remote web-based human machine interface. However, this algorithm has high average end-to-end delay.

A novel segmentation algorithm [10] was proposed for separating almond nut and shell from locale and outline. A novel artificial neural network algorithm combined with image processing technique was developed. At first, the appropriate group of color features was extracted from the images. The extracted features were selected by using sensitivity analysis. Finally, the selected features were classified based on the artificial neural network, which was used to classify into three classes such as object, outline and locale. The segmentation was improved by removing the effect of the outline around the objects. However, this ANN based classification algorithm has high training time.

The morphology modelling in the CIE L*a*b* color space [11] was developed for segmenting the plants from the images. The morphology modelling was applied in supervised learning phase for dealing with the color features of the crop in order to pixel lightness element and for achieving the crop color model. In addition, the performance of the segmentation of cloudy, overcast and sunny day's images was evaluated. However, the segmentation performance was not significantly improved for different types of structuring features.

A model based method [12] was proposed to recognize the leaves from natural images for identifying the tree species. Initially, the leaf images in the complex natural locale were

collected. The contour of the leaf was retrieved by using two-step active contour segmentation algorithm according to the polygonal leaf representation. After, the features utilized were high-level geometrical descriptors which provide the possible semantic interpretation. These geometrical descriptors have better performance than the generic and statistical shape descriptors. However, this technique has high processing time and more complicated for large database.

The image processing techniques [13] was proposed for detecting plant leaves diseases. The image processing technique involves different processes such as image acquisition, image pre-processing, feature extraction and classification. Initially, the different leaf images were collected with required resolution. The acquired images were pre-processed for enhancing features and removing undesired information. Then the image was segmented using segmentation algorithms and features were extracted through texture based feature extraction techniques. Finally, the extracted features were classified using neural network classifiers to classify the plant leaf disease. However, neural network classifiers were complexity to know about how many neurons and layers are needed.

The leaf disease detection method [14] was proposed for Cercospora leaf spot in sugar beet by using a template matching algorithm. The orientation code matching algorithm and site-specific observations of the disease improvement in sugar beet plants were described. This robust template matching algorithm was developed for non-rigid plant object searching based on illumination, translation, slight rotation and occlusion modification. Moreover, single-feature two-dimensional XY-color histogram was introduced and support vector machine classifier was employed for pixel-wise disease classification and quantization. However, the capturing of all types of illumination under different weather conditions was insufficient for the classification process.

The monitoring system [15] was proposed for crop production by using the clustering approach. The sub-pixel cover fraction and unconstrained spectral signature of the plant feature from the combined hyper spectral signal was extracted simultaneously by developing the multiple endmember spectral mixture analysis. The generation of lookup tables (LUT) was achieved through a radiative transfer model for both soil and plant features. After segmenting the features, the clustering method was performed for enhancing the efficiency of utilization of LUT in this proposed model. Then, the Bayesian selection algorithm was presented for selecting the most favourable clusters. However, the influences of outline in mixture and non-linear mixed effects in orchard systems were not considered.

MATERIALS AND METHODS

Improved Monitoring System for Crop Production

Our proposed monitoring system for improving crop yield monitoring includes three different phases such as image pre-processing, detection phase and communication phase. In each phase, different data mining approaches are developed for different processes. The first phase focuses on the pre-processing of soil and plant images. The second phase performs the prediction process of plant disease and its causes due to the soil features. The final phase employs the information technology to alert the farmers by distributing the types of plants, diseases of plants and its causes. In this section, these three different phases in monitoring system are explained briefly.

Image Pre-processing Phase

Initially, the input images of different plants and soil are collected by using the digital camera with a desired resolution for better quality. The captured images are transmitted through either wired or wireless network to the image processing unit for further processing. Then the collected images are pre-processed for removing the noise or other disturbances from the images and enhancing the features for prediction process. The noise from the images is eliminated and RGB images are converted into the color space representations. Then, the enhanced images are segmented by using Region of Interest (ROI) based segmentation method. ROI segmentation is used to segment the certain images into the number of regions or classes. It is denoted as follows:

$$R_{i\geq j} = \frac{\psi.\frac{|R_i||R_j|}{|R_i|+|R_j|}.\|u_i - u_j\|^2 + (1-\psi).e\omega}{1\left(\partial\left(R_i, R_j\right)\right)} \quad (1)$$

In equation (1), R_i and R_j refer the two adjacent regions of the area, $\partial\left(R_i, R_j\right)$ is the length of the two regions, ω refers the weight of $(0,1)$, e denotes the boundary strength and u_i, u_j represent the spectral values of two regions.

The contour features of plant images are retrieved through the polygonal leaf model. Then, semantic representation is applied for these contour features for identifying the types of plant with the help of botanical plant species dictionary in which description of different types of plant species are stored.

Then, the features are extracted from plant images, based on the texture, color and shape feature extraction methods. The shape features from the plant images are extracted using the inner-distance which is utilized for constructing the shape descriptors based on shape contexts. Inner-distance is the length of the shortest route between the sight points within the shape boundary. The shape O is defined as connected and closed subset of R^2. The inner distance between the two

points (x_i, x_j) is represented as $d(x_i, x_j; O)$. The shape context at point x_i is defined as the histogram h_i of the relative coordinates of the remaining n-1 points.

$$h_i(k) = \#\{x_j : j \neq i, x_j - x_i \in bin(k)\} \quad (2)$$

The similarity between the two points is estimated as weighted combination and is given as,

$$D = aD_{ac} + D_{sc} + bD_{be} \quad (3)$$

In the above equation (3), D_{ac} refers the appearance difference, D_{be} refers the bending energy and D_{sc} denotes the shape context distance which is an average distance between two points. Here $a\ and\ b$ are weights. The shape context utilizes the inner-distance for determining the spatial relation between sight points. Hence, the inner-angle $\theta(x_i, x_j; O)$ is provided for the orientation bins. The geometric parameters are extracted based on the fractal geometry. The geometrical object X can be measured its length N using rule with length u. Then the fractal dimension is measured as,

$$D_x \propto \lim_{u \to 0} \frac{\log N}{\log u} \quad (4)$$

Plant Disease Prediction

According to the plant types and plant features, the plant disease is predicted by the Transductive Support Vector Machine (TSVM). TSVM is used for discovering the presence of disease in plants compared with training dataset. Assume training features of plant images be $(x_1, y_1), \dots (x_l, y_l), y \in \{-1,1\}$ and $x_1^*, \dots, x_k^*$ be non-labelled plant image features. The TSVM is defined as the following optimization problem.

$$\min_{y^*, w, b, \xi, \xi^*} \frac{1}{2}\|w\| + C\sum_{i=1}^{n} \xi_i + C^* \sum_{j=1}^{k} \xi_i^* \quad (5)$$

Subject to, $y_i\big((w.x_i) + b\big) \geq 1 - \xi_i \ \ i = 1, \dots, n$

$$y_j\left(\left(w.x_j^*\right) + b\right) \geq 1 - \xi_j^* \ \ j = 1, \dots, k$$

$$\xi_i \geq 0, \qquad i = 1, \dots, n$$

$$\xi_j^* \geq 0, \qquad j = 1, \dots, k$$

This optimization problem is solved by the training process of TSVM. The training algorithm has containing the following stepsheadings.

1. Set parameter C and C^*
2. Utilize inductive SVM on training data to obtain initial classifier
3. Set the number of positive label samples based on rule
4. Compute decision function values for all unlabelled samples
5. Label samples with the highest decision function as positive
6. Set temporary effect factor C_{tmp}^*
7. Retrain the support vector machine over entire samples
8. Switch labels of one pair of different-labelled unlabelled samples using certain rule to make the value of objective function by using (5)
9. Repeat the process until no pair of samples satisfying the switching condition
10. Increase the value of C_{tmp}^* then move to step 7
11. $If (C_{tmp}^* > C^*)$
12. Stop
13. Obtain the output

In testing phase, the data is compared with training data and the type of plant disease is predicted effectively. In addition, the required water content level for the specific type of plant is predicted based on the features extracted from the soil images by using LDA technique.

Prediction of Cause for Plant Disease

Moreover, the texture and color features are extracted from the soil images for predicting the causes for plant leaf diseases. The texture features are such as contrast, intensity, and correlation. The color features are hue, saturation, and value or RGB color spaces. The LDA is used for modeling the features of soil images with the descriptions of the plant diseases which measures the probability between soil and plant features.

For every image indexed by $m \in \{1, \dots, M\}$ in a plant species dictionary:

1. Select K-dimensional topic weight vector θ_m from Dirichlet distribution, $p(\theta|\alpha) = Dirichlet\ (\alpha)$
2. For every patch indexed by $n \in \{1, \dots, N\}$ in the image:
3. Select an object category $z_n \in \{1, \dots, K\}$ from the multinomial distribution, $p(z_n = k|\theta_m) = \theta_m^k$
4. Given the selected object category z_n, draw a codeword x_n from the probability $p(x_n = i|z_n = j, \beta) = \beta_{ij}$

The computed probability values are given to the ANN classification for predicting whether the soil features are causes of plant diseases or not. The probability values are given to train the ANN. ANN has three layers namely input, hidden and output layer. The probabilities are denoted as

$f(x) = x$ are given to the input layer of neurons. The hidden layer of ANN is defined as tan-sigmoid transfer function.

$$f(x) = \frac{2}{1 + e^{-2x}} - 1 \qquad (6)$$

The output layer of ANN is described by the following equation.

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \qquad (7)$$

In the above equation (6), y is the output neuron value; f(x) is the transfer function, w_i is the weight values, x_i is input data values and b refers to the bias value. The output neuron value is the soil features that are the causes for that plant disease.

Communication Phase

In the communication phase, the predicted plant disease type information and its causes due to soil are passed to the farmers in terms of text messages or emails including some potential suggestions such as increasing manure, irrigation and etc. Thus this proposed monitoring system can improve the productivity of farmers through effective irrigation and disease prevention.

Algorithm

Input: Plant or Soil images $I_1, \ldots, I_N$

Output: Type of plant & disease, and its causes.

1. Collect images
2. Remove noise from image
3. Convert image into color space representation

//Image segmentation

4. Select seed pixel in image
5. Assign criteria to grow the region
6. Include pixel in the region if it is 8-connected to at least one of the pixel in the region
7. Test all the pixels for allocation
8. Label all the regions
9. Combine regions if two regions have the same label
10. Extract the shape, geometric features from the plant images
11. Apply semantic representation for contour
12. Predict the type of plant

//Prediction of plant disease

13. Define optimization problem for TSVM classifier
14. Obtain training samples
15. Perform TSVM process
16. Obtain result
17. Compare training data with testing data
18. Find the type of disease

//Prediction of the causes of plant diseases

19. Extract texture and color features from the soil images
20. Select the topic mixture for the plant species description based on Dirichlet distribution
21. Generate each word in the description by selecting the topic
22. Then generate word using topic based on multinomial distribution
23. Compute the probability between soil features and words
24. Perform ANN classification using the computed probabilities
25. Obtain the output function by using the equation (7)
26. Predict the causes for the particular plant disease
27. Inform the prediction results to farmers through text messages
28. End

RESULTS AND DISCUSSION

In this section, the performance of both prediction of plant diseases and its causes is evaluated for existing and proposed techniques in terms of precision and accuracy. The results of prediction of causes for plant diseases are compared to the similarity measure based technique and LDA based technique. The results of plant disease prediction are compared to the SVM and TSVM technique.

Precision

Precision value is defined as the ratio of predicted features that are relevant and evaluated at the true positive outcomes. It is computed as,

$$Precision = \frac{Truepositive(TP)}{Truepositive(TP) + Falsepositive(FP)}$$

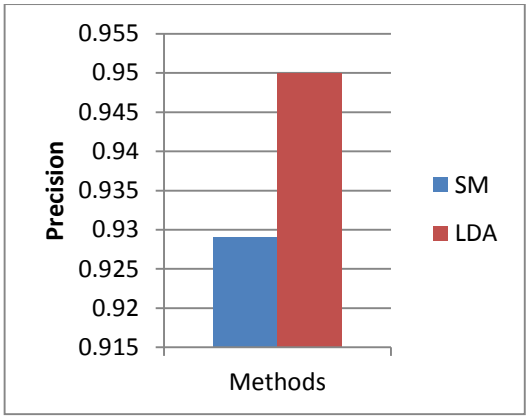

Figure 1: Comparison of Precision for Prediction of Plant Disease Causes

Figure 1 shows that the comparison of SM and LDA techniques in terms of precision for prediction of causes for plant diseases. The analysis from the above graph proves that the precision of LDA technique increases with increasing the number of features. In the x-axis, methods are taken and in the y-axis the precision values are taken. It shows that the precision is increased in LDA prediction method.

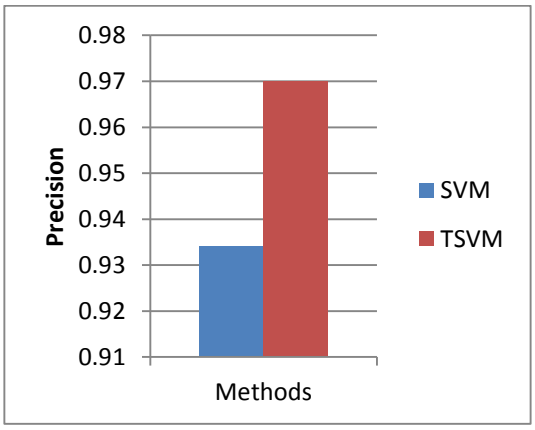

Figure 2: Comparison of Precision for Plant Leaf Disease Prediction

Figure 2 shows that the comparison of SVM and TSVM techniques in terms of precision for the plant leaf disease prediction. The analysis from the above graph proves that the precision of TSVM technique increases with increasing the number of features. In the x-axis, methods are taken and in the y-axis the precision values are taken. It shows that the precision is increased in TSVM prediction method.

Accuracy

Accuracy is defined as the ratio of true positives and true negatives to the sum amount of features examined. It is measured as,

$$Acc = \frac{\text{TP} + \text{True negative}}{\text{TP} + \text{True negative} + \text{FP} + \text{False negative}}$$

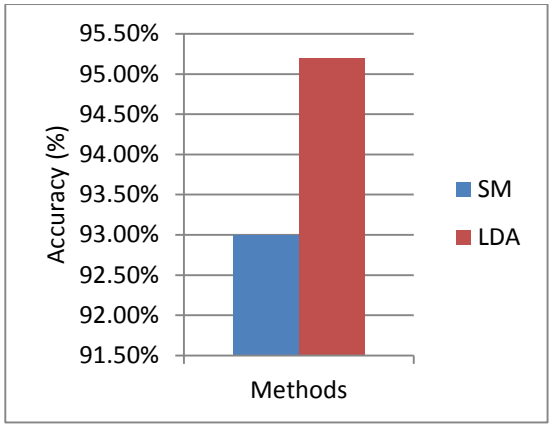

Figure 3: Comparison of Accuracy (%) for Prediction of Plant Disease Causes

Figure 3 shows that the comparison of SM and LDA techniques in terms of accuracy for prediction of causes for plant diseases. The analysis from the above graph proves that the accuracy of LDA technique increases with increasing the number of features. In the x-axis, methods are taken and in the y-axis the accuracy (%) is taken. It shows that the accuracy is increased in the LDA prediction method.

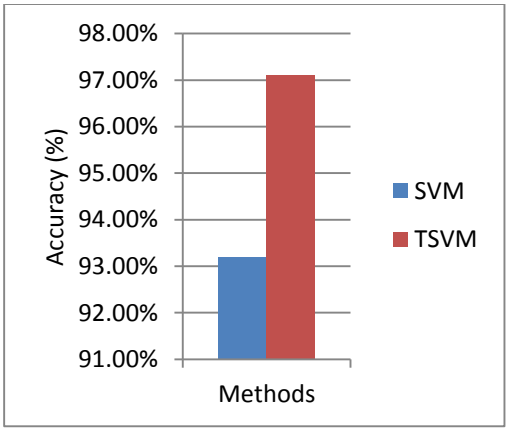

Figure 4: Comparison of Accuracy (%) for Plant Leaf Disease Prediction

Figure 4 shows that the comparison of SVM and TSVM techniques in terms of accuracy for plant leaf disease prediction. The analysis from the above graph proves that the accuracy of TSVM technique increases with increasing the number of features. In the x-axis, methods are taken and in the y-axis the accuracy (%) is taken. It shows that the accuracy is increased in TSVM prediction method.

CONCLUSION

In this article, a novel data mining technique for monitoring systems in agricultural fields is proposed. The proposed data mining technique is related to the prediction of plant diseases and its causes for plant diseases. In this proposed system, the type of plant is identified from the extracted features of leaf images by contour using botanical plant species dictionary. Then, the plant disease is predicted by TSVM classification by shape and texture features of the plant images. The soil image features are modeled with diseased plant images based on the LDA technique with help of color and texture features of soil images. Moreover, the causes for the plant diseases are predicted by using ANN classification. Then the predicted information's are sent to the farmers in terms of text messages through mobile phones. Thus, our proposed monitoring system is used for improving the crop yield monitoring system in terms of proper irrigation and disease control. The experimental results prove the effectiveness of the proposed monitoring system. In future, the prediction technique for monitoring the growth level of the plant will be taken into consideration in order to improve the crop productivity.

REFERENCES

[1] Mishra, S., Mishra, D., and Santra, G. H. "Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper", .Indian Journal of Science and Technology, 9(38), 2016.

[2] Milovic, B., and Radojevic, V. "Application of data mining in agriculture", Bulgarian Journal of Agricultural Science, 21(1), pp. 26-34, 2015.

[3] Ramesh, D., and Vardhan, B. V. "Data mining techniques and applications to agricultural yield data," International Journal of Advanced Research in Computer and Communication Engineering, 2(9), pp. 3477-80, 2013.

[4] Huang, W., Guan, Q., Luo, J., Zhang, J., Zhao, J., Liang, D., ... and Zhang, D. "New optimized spectral indices for identifying and monitoring winter wheat diseases," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7(6), pp. 2516-2524, 2014.

[5] Bolten, J. D., Crow, W. T., Zhan, X., Jackson, T. J., and Reynolds, C. A. "Evaluating the utility of remotely sensed soil moisture retrievals for operational agricultural drought monitoring," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 3(1), pp. 57-66, 2010.

[6] Barbedo, J. G. A. "A novel algorithm for semi-automatic segmentation of plant leaf disease symptoms using digital image processing," Tropical Plant Pathology, 41(4), pp. 210-224, 2016.

[7] Han, L., Haleem, M. S., and Taylor, M. "A novel computer vision-based approach to automatic detection and severity assessment of crop diseases," In Conference on Science and Information (SAI), pp. 638-644, 2015.

[8] Xiaoxia, Y., and Chengming, Z. "A soil moisture prediction algorithm base on improved BP," In Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics), pp. 1-6, 2016.

[9] Sung, W. T., Chung, H. Y., and Chang, K. Y. "Agricultural monitoring system based on ant colony algorithm with centre data aggregation" IET Communications, 8(7), pp. 1132-1140, 2014.

[10] Teimouri, N., Omid, M., Mollazade, K., and Rajabipour, A. "A novel artificial neural networks assisted segmentation algorithm for discriminating almond nut and shell from background and shadow," Computers and Electronics in Agriculture, 105, pp. 34-43, 2014.

[11] Bai, X. D., Cao, Z. G., Wang, Y., Yu, Z. H., Zhang, X. F., and Li, C. N. "Crop segmentation from images by morphology modeling in the CIE L* a* b* color space," Computers and electronics in agriculture, 99, pp. 21-34, 2013.

[12] Cerutti, G., Tougne, L., Mille, J., Vacavant, A., and Coquin, D. Understanding leaves in natural images–a model-based approach for tree species identification. Computer Vision and Image Understanding, 117(10), pp. 1482-1501, 2013.

[13] Gavhale, K. R., and Gawande, U. "An Overview of the Research on Plant Leaves Disease detection using Image Processing Techniques" IOSR Journal of Computer Engineering, 2014.

[14] Zhou, R., Kaneko, S. I., Tanaka, F., Kayamori, M., and Shimizu, M. "Disease detection of Cercospora Leaf Spot in sugar beet by robust template matching" Computers and Electronics in Agriculture, 108, pp. 58-70, 2014.

[15] Tits, L., Somers, B., and Coppin, P. "The potential and limitations of a clustering approach for the improved efficiency of multiple endmember spectral mixture analysis in plant production system monitoring," IEEE Transactions on Geoscience and Remote Sensing, 50(6), pp. 2273-2286, 2012.