IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 7, Issue.11, Nov 2019 E-ISSN: 2347-2693

Adoptive Clustering Algorithm with Feature Subset Selection Method to find the Plant Diseases

Anuradha Anumolu[1*], Shaheda Akthar[2]

[1]Research Scholar, Acharya Nagarjuna University, Guntur Dist.
[2]Lecturer in Computer Science Government College for Women (A) & Research Supervisor. Dept of Computer Science and Engineering, Acharya Nagarjuna University, Guntur Dist.

DOI: https://doi.org/10.26438/ijcse/v7i11.198202 | Available online at: www.ijcseonline.org

Accepted: 15/Nov/2019, Published: 30/Nov/2019

***Abstract*:** Machine Learning (ML) is the subfield in Artificial Intelligence (AI) that works dynamically to solve several issues. ML mainly focused on understanding the structure of the data and selecting the specific model based on the given dataset. Nowadays plant diseases are becoming very dangerous to farmers. Various plant diseases are identified by many researchers based on the pathogen. Several visible and invisible features are present to identify plant diseases. Visible features such as shape, size, silting are most widely used to analyze the condition of the plant. In this paper, the adaptive clustering algorithm (ACA) is introduced to detect diseases in plants. To show the disease-affected region the fuzzy c-means (FCM) clustering approach is adopted to highlight the disease-affected region with red patches which are called clusters. To improve the performance of the proposed approach the feature subset selection is used to increase the effectiveness and scalability. The output results show the performance of the ACA.

***Keywords*:** Machine learning (ML), ACA, AI and K-Means.

I. INTRODUCTION

Plant diseases cause more problems for the production of agriculture. It is very important to find the plant diseases in the early stages because this will show the impact on crop production [1]. Early detection of diseases is useful to control and prevent plant diseases that play a significant role in managing and making decisions in agriculture production. For the past many years, finding plant diseases is a crucial task.

Plants that effected with diseases shows the marks or lesions on leaves, stems, flowers, or fruits. In general, every disease has its own patterns and properties to show the condition of the disease based on the abnormalities. For every plant, to detect the disease several basic symptoms that are visible or invisible in the leaves [2]. Several diseases that are frequently occurring regularly may cause the economic loss. It is very important to detect the plant disease on time. Clustering is most widely used in many applications. It is used to detect the similar objects in any data. This will do the partition the set of objects that are assigned according to class labels.

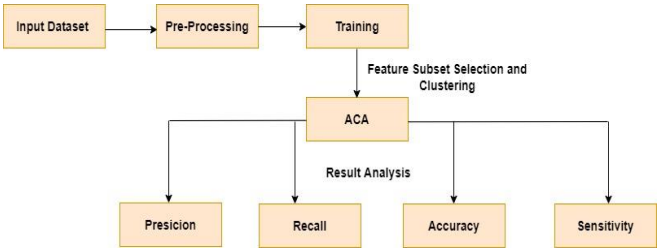

Figure: 1 Step for Plant Disease Detection and Clustering

Manually, there are many machine learning approaches are developed to prevent plant diseases. To improve the performance of the proposed approach the feature subset selection (FSS) is integrated to increase the performance of disease prediction. FSS is the dimensionality reduction approach that extracts the efficient features that shows the impact on disease prediction. In this paper, clustering plays a significant role in detecting the diseases by showing the patches on the leaves. The adaptive clustering approach (ACA) is used to form accurate clusters to detect the disease-affected patches that are in red.

II. LITERATURE SURVEY

L. Li et al., [4] discussed about various deep learning approaches that are used to detect the disease in crop leaf. The author explained about several image improving techniques and deep learning approaches. These are most

widely used to detect the plant diseases and insect pests. X. Liu et al., [5] introduced the unique proposed approach which is integrated with the LSTM and feature extraction technique that extracts the patch features from the network. The proposed approach is applied on public plant dataset that consists of 271 plant diseases consists of 220,592 images. The proposed approach shows the improvement in detection plant diseases. Jayme et al., [6] presented the observation of new significant factors that are mainly affects the design of deep learning approaches that are applied to plant pathology. The proposed approach focused on in-depth analysis that shows the advantages and disadvantages about the existing and proposed approaches. For experimental results the public available dataset is used to show the performance of proposed approach. This dataset consists of 50k images that are various types of images. P. Shah et al., [7] presented several image processing and ML approaches that are used to diagnose the plant disease detection and improved classification. Based on the types of plant disease the detection rate is shown. These approaches are mainly following the total no of classes, pre-processing techniques, segmentation techniques etc. These approaches are used to propose the detection and classification of rice plants diseases. L. Shanmugam et al., [8] introduced the automated approach that is used to detect the plant diseases using remote sensing images. This is also used to detect the plant diseases in the early stages. This approach consists of two phases such as training and testing. In the training stage, several features are extracted based on the threshold values from the image and in the second stage focused on monitoring the crops and finding the specific diseases by using canny edge detection. Q. Wu et al., [9] introduced the new deep convolution generative adversarial networks (DCGAN) which is achieved the accuracy of 94.33%. By using this approach an effective detection of plant diseases are done. This approach is mainly focused on disease detection in tomato leaf. This system is applied on large tomato leaf datasets and a better training is also given in this.

M.A.H.Abas et al., [10] discussed about the using of VGG16 for better classifications of plants. Various flower images are used to recognise the shape and structure of the crop leaves that are similar in nature. In this paper, the data augmentation approach is used to reduce the over fitting issues in CNN which is applied to little amount of data. The proposed approach uses the VGG16 for training and this consists of 2800 flower images. The proposed approach achieves the 96.25% accuracy training and 93.93% for validation and 89.96% for testing. P. Jiang et al., [11] proposed the approach called as new apple leaf disease detection approach that utilizes the deep CNN and this is used the GoogLeNet Inception structure and Rainbow merging. The proposed approach INAR-SSD is trained to detect the five general apple leaf diseases. From the evaluation results the proposed approach achieved the 78.80% accuracy. B. Wang et al., [12] proposed the classification approach that uses the k-nearest neighbor (kNN) classifier and spatial structure optimizer (SSO) to detect the plant diseases. For experimental analysis they have used the Flavia, Swedish and Leafsnap datasets to know the performance. Results show the high performance of image classification. W. Yang et al., [13] introduced the DL approach that extract the spectral features from the given image datasets. This mainly used to detect the clod damage in corn seeds.

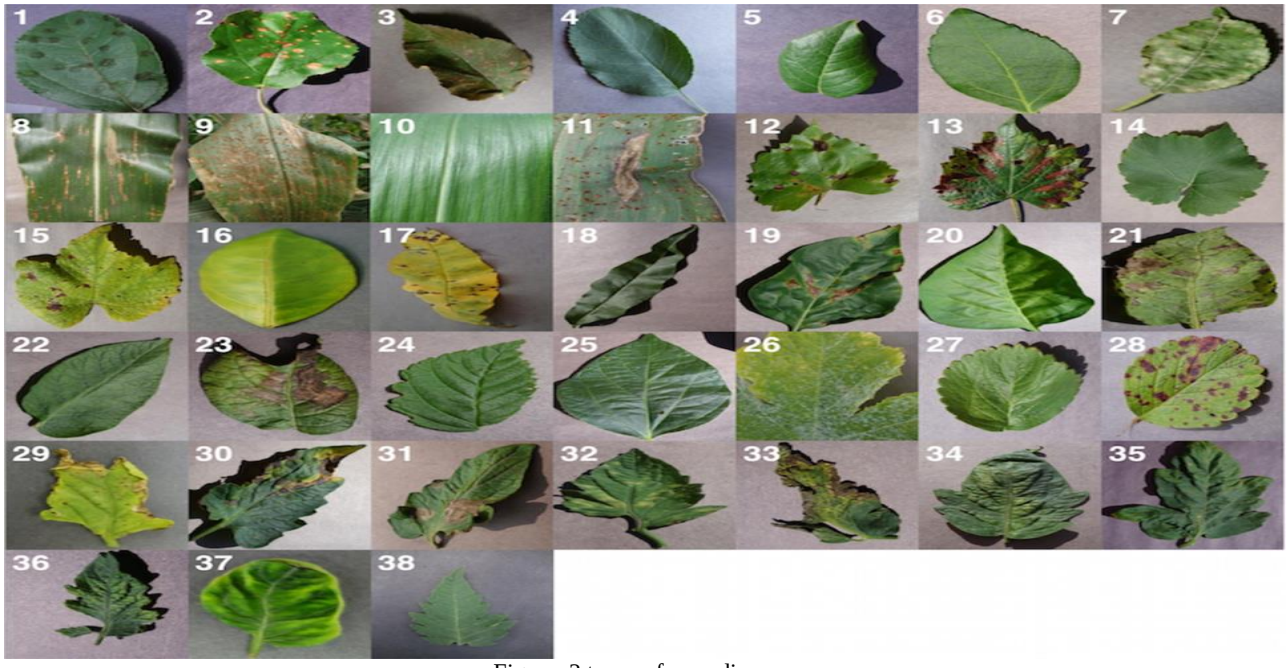

Figure: 2 types of crop diseases

Dataset Description

The dataset is collected from kaggle website which is called as Plant Village dataset. This dataset consists of 54,306 images belongs to plant leaves that are having 38 class labels. Every label consists of crop-disease pair which is used to predict the crop-disease pair gives the plant leaf. The size of the images is 256x256 pixels to perform the model optimization and predictions on these downscaled images.

Feature subset selection (FSS)

FSS is the approach used to remove the features that are not required or not relevant. The subset selects the features by following the Occam's razor principle which gives the better performance according to the objective method. Sometimes this is called as NP-hard (nondeterministic polynomial-time hard) problem [14] [15]. The sizes of the features are increased from the past 5 years and this is one the significant feature selection approach which is used for the better classification. Compare with feature extraction methods the feature selection approaches represents the actual data. The aim of the FSS is to prevent the over fitting to improve the performance. The features such as size of the leaf, affected leaf area, size of the image, shape of the input image, type of patches, colour of the images etc.

i. In this stage, the filters are used to extract the better features from the input data without any training involved.

ii. Wrappers utilized the learning approaches to analyze the most useful features.

iii. Several embedded approaches are merged with feature selection step to develop the classifier.

Adoptive clustering Algorithm with feature subset selection method (ACA-FSS)

- This algorithm has been implemented above using bottom up approach. It is also possible to follow top-down approach starting with all data points assigned in the same cluster and recursively performing splits till each data point is assigned a separate cluster.
- The decision of merging two clusters is taken on the basis of closeness of these clusters. There are multiple metrics for deciding the closeness of two clusters:

$$\textbf{Euclidean distance: } \|a\text{-}b\|_2 = \sqrt{\Sigma(a_i\text{-}b_i)}$$

- This algorithm is integration of fuzzy c-means algorithm and the following steps are as follows.

Let X = $\{x_1, x_2, x_3 ..., x_n\}$ be the set of data points and V = $\{v_1, v_2, v_3 ..., v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) calculate the fuzzy membership $'\mu_{ij}'$ using:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}(\frac{d_{ij}}{d_{ik}})^{(2lm-1)}}$$

3) compute the fuzzy centres $'v_i'$ using:

$$V_j = (\sum_{i=1}^{n}(\mu_{ij})^m x_i) / (\sum_{i=1}^{n}(\mu_{ij})^m), \forall j = 1,2, \dots \dots c$$

4) Repeat step 2) and 3) until the minimum *'J'* value is achieved or $||U^{(k+1)} - U^{(k)}|| < \beta$.

where,

'k' is the iteration step.

'β' is the termination criterion between [0, 1].

'$U = (\mu_{ij})_{n*c}$' is the fuzzy membership matrix.

'J' is the objective function.

III. EXPERIMENTAL RESULTS

The implementation is done with java with net beans 8.0.2 and jdk 1.8. The dataset contains 54,306 images. For training 10,000 images are used and for testing 15,000 images are used and detected various diseases affected plants are shown by using clustering.

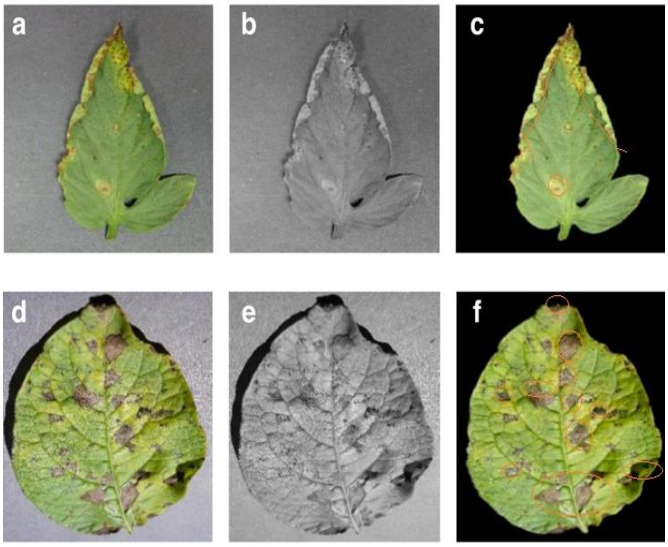

Figure: 3 (a), (d) are inputs and (b) (e) are the gray images and (c) (f) are the clustering outputs.

Performance Metrics

The performance of the proposed system is evaluated by showing the metrics such as sensitivity, specificity and accuracy. The count values are such as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

Precision

The proportion of actual positives which are correctly identified is the measure of the sensitivity. It relates to the ability of the test to identify positive results.

$$\text{Presicion} = \frac{\text{No. of TP}}{\text{No. of TP} + \text{No. of TN}}$$

Specificity

The proportion of negatives which are correctly identified is the measure of the specificity. It relates to the ability of the test to identify negative results.

$$\text{Specificity} = \frac{\text{No. of TN}}{\text{No. of TN} + \text{No. of FP}}$$

Accuracy: This will calculate the overall accuracy of the images classified.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Recall: Appropriate when **minimizing false negatives** is the focus.

$$\textbf{Recall} = \frac{\text{TP}}{\text{No. of TP} + \text{No. of FN}}$$

F1 Measure

$$\text{F1 Measure} = 2 \times \frac{\text{accuracy} * \text{recall}}{\text{accuracy} + \text{recall}}$$

Clustering percentage: The overall percentage of clustering is represented as

$$C_j = \text{Cluster}(X_i) = \arg_j \min || X_i - \mu_j ||^2$$

$$\text{percentage of Cluster} = \sum_{i=1}^{m} (x_i - c_i)^2 = \sum_{j=1}^{k} \sum_{i \in \text{OwnedBy}(\mu_j)} (X_i - \mu_j)^2$$

(within cluster sum of squares)

Table: 1 shows the % of affected area and clustering percentage

Algorithms	Disease Affected Area(accuracy)	Precision	Recall	Specificity	Percentage of Clustering (%)
K-Means Clustering	65%	67.9%	69.09%	69.12 %	59%
Adaptive Clustering (ACA)	87%	88.12 %	88.45%	89.12 %	85.6%

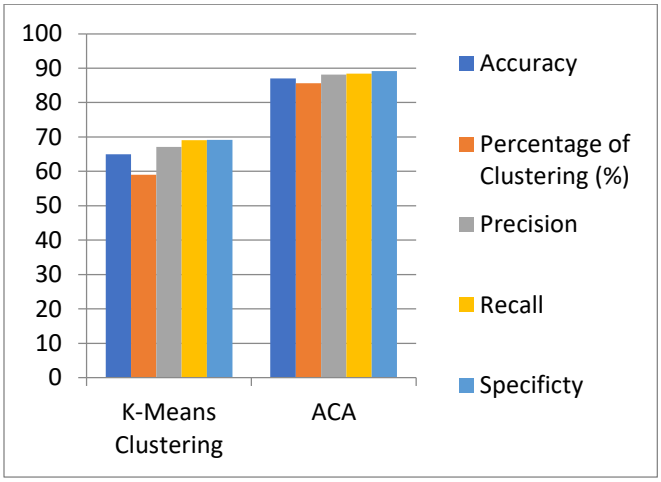

Figure 4 Comparative Performance

Table 1 shows the performance of the proposed clustering algorithm. The performance id increased based on the formation of the clusters and quality of the clusters. In the selected datasets, there are number of diseases effected plant images that are present. The proposed ACA is integration of fuzzy c-means clustering (FCM) and FSS which calculates the abnormal area within the selected image.

IV. CONCLUSION

In this paper, the ACA-FSS is implemented to find the disease with the help of clustering. Clustering is used to find the red patches that affected in the plant leaf and finds the severity of the diseases. By using the feature subset selection the features are extracted. These features show the huge impact on results. The experimental results are shown in table 1. The proposed ACA is significantly improves the result in terms of disease affected area and percentage of clustering which is used to identify the disease of plants.

REFERENCES

[1] F. Fina, P. Birch, R. Young, J. Obu, B. Faithpraise, and C. Chatwin, "Automatic plant pest detection and recognition using k-means clustering algorithm and correspondence filters," Int. J. Adv. Biotechnol. Res., **vol. 4,no. 2, pp. 189–199, Jul. 2013.**

[2] M. A. Ebrahimi, M. H. Khoshtaghaza, S. Minaei, and B. Jamshidi, "Vision-based pest detection based on SVM classification method," Comput. Electron. Agricult., **vol. 137, pp. 52–58, May 2017.**

[3] S. R. Dubey and A. S. Jalal, "Adapted approach for fruit disease identification using images," Int. J. Comput. Vis. Image Process., **vol. 2, no. 3, pp. 44–58, Jul. 2012.**

[4] L. Li, S. Zhang and B. Wang, "Plant Disease Detection and Classification by Deep Learning—A Review," in IEEE Access, **vol. 9, pp. 56683-56698, 2021**, doi: 10.1109/ACCESS.2021.3069646.

[5] X. Liu, W. Min, S. Mei, L. Wang and S. Jiang, "Plant Disease Recognition: A Large-Scale Benchmark Dataset and a Visual Region and Loss Reweighting Approach," in IEEE Transactions on Image Processing, **vol. 30, pp. 2003-2015, 2021,** doi: 10.1109/TIP.2021.3049334.

[6] Jayme G.A. Barbedo, Factors influencing the use of deep learning for plant disease recognition, Biosystems Engineering, **Volume 172, 2018**, https://doi.org/10.1016/j.biosystemseng.2018.05.013.

[7] P. Shah, H. B. Prajapati and V. K. Dabhi, "A survey on detection and classification of rice plant diseases," 2016 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC), **pp. 1-8, 2016.** doi: 10.1109/ICCTAC.2016.7567333.

[8] L. Shanmugam, A. L. A. Adline, N. Aishwarya and G. Krithika, "Disease detection in crops using remote sensing images," 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), **pp. 112-115, 2017.** doi: 10.1109/TIAR.2017.8273696.

[9] Q. Wu, Y. Chen, and J. Meng, "DCGAN-based data augmentation for tomato leaf disease identification," IEEE Access, **vol. 8, pp. 98716–98728, 2020.**

[10] M. A. H. Abas, N. Ismail, A. I. M. Yassin and M. N. Taib, "VGG16 for plant image classification with transfer learning and data augmentation", Int. J. Eng. Technol., **vol. 7, pp. 90-94, Oct. 2018.**

[11] P. Jiang, Y. Chen, B. Liu, D. He, and C. Liang, "Real-time detection of apple leaf diseases using deep learning approach based on improved convolutional neural networks," IEEE Access, **vol. 7, pp. 59069–59080, 2019.**

[12] B. Wang and D. Wang, "Plant leaves classification: A few-shot learning method based on Siamese network," IEEE Access, **vol. 7, pp. 151754–151763, 2019.**

[13] W. Yang, C. Yang, Z. Hao, C. Xie, and M. Li, "Diagnosis of plant cold damage based on hyperspectral imaging and convolutional neural network," IEEE Access, vol. 7, pp. 118239–118248, 2019.

[14] tansal, P., Yadav, H. and Sunkaria, R.K., "Impulse noise removal usinm MDBUTMF gith histomram estimation", In Advance Computing Conference (IACC), International on, **pp. 468-471, IEEE, 2015.**

[15] Ramakrishnan M., Sahaya Anselin Nisha, "Groundnut leaf disease detection and classification by usinm back probamation almorithm", In Communications and Signal Processing (ICCSP), International Conference on, **pp. 0964-0968, IEEE, 2015.**

