

Improving prediction of plant disease using k-efficient clustering and classification algorithms

Asraa Safaa Ahmed¹, Zainab Kadhm Obeas², Batool Abd Alhade², Refed Adnan Jaleel³

¹Department of Computer Sciences, College of Science, Diyala University, Diyala, Iraq

²College of Science, Al-Qasim Green University, Babylon, Iraq

³Department of Information and Communications Engineering, Information Engineering College, Al-Nahrain University, Baghdad, Iraq

Article Info

Article history:

Received Sep 22, 2021

Revised Apr 7, 2022

Accepted May 6, 2022

Keywords:

K-means

K-medoids

K-nearest neighbors

Plant disease

Prediction

Soybean

ABSTRACT

Because plant disease is main cause of most plants' damage, improving prediction plans for early detection of plant where it has disease or not is an essential interest of decision makers in the agricultural sector for providing proper plant care at appropriate time. Clustering and classification algorithms have proven effective in early detection of plant disease. Making clusters of plants with similar features is an excellent strategy for analyzing features and providing an overview of care quality provided to similar plants. Thus, in this article, we present an artificial intelligence (AI) model based on k-nearest neighbors (k-NN) classifier and k-efficient clustering that integrates k-means with k-medoids to take advantage of both k-means and k-medoids to improve plant disease prediction strategies. Objectives of this article are to determine performance of k-mean, k-medoids and k-efficient also we compare k-NN before clustering and with clustering in prediction of soybean disease for selecting best one for plant disease forecasting. These objectives enable us to analysis data of plant that help to understand nature of plant. Results indicate that k-NN with k-efficient is more efficient than other in terms of inter-class, intra-class, normal mutual information (NMI), accuracy, precision, recall, F-measure, and running time.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Refed Adnan Jaleel

Department of Information and Communication Engineering, Information Engineering College

Al-Nahrain University, Baghdad, Iraq

Email: Iraq_it_2010@yahoo.com

1. INTRODUCTION

Detection of Plant disease is an important aspect of precision agriculture since it focuses on detecting diseases in their early stages [1]. Agriculture is the sort of productivity that the world needs to develop. Despite the fact that there are several projects to expand each sector, there has been little progress in the agriculture farming sector. The application for the detection of plants is therefore to enhance the agricultural sector and to increase safety criteria so that the development of the plant growth can be developed [2].

Plant monitoring is one of the greatest frameworks that agriculture sector institutions should build because of the spread of plant disease [3]. Also, early detection of plant disease is the greatest approach to enhance prediction, and this is where machine learning (ML) algorithms may help in the current work. The integration of clustering with classification facilitates the use of very smart frameworks which take account of the operational performance and efficiency of all ML-affiliated institutions [4], [5] and detect knowledge patterns that correlated to the nature of plant diseases. ML can address critical concerns connected to plant

disease if information about the plants is obtained. It's utilized to derive diagnostic rules and give decision-makers a precise prediction method [6].

In fact, thanks to clustering and classification algorithms, creating clusters of plants with similar characteristics allows for a better understanding of the quality of care given to various plants and by using classifiers it is possible to build a forecasting model of a plant disease by examining the historical data. This article was conducted to understand better the most novel and practical applications of integrating clustering and classification algorithms of ML in the soybean plant disease prediction. Soybeans are an useful crop since it's processed for their oil and meal [7]. People eat meat and oil from the meal, which is then given to animals that are commonly eaten by humans. Different diseases damage soybean crops [8].

Recently, several researchers have studied the application of the integration of clustering and classification algorithms into plant disease prediction as a response to the expansion of the plant disease pandemic. The research question is being followed. Research question (RQ): How does combining clustering and classification help predict and control a pandemic of plant diseases? Based on this research question, we collect articles from different good scientific databases. The key search terms were ML, clustering, classification, data mining, k-medoids, k-means, plant disease, and soybean. Boolean operations such as OR and AND used literary search in the many databases to provide important results for the research articles concerned.

Kaushal and Bala [9] proposed using k-mean for the segmentation of input images and support vector machine (SVM) classifier was applied for classifying the input image into two classes and they enhanced performance of SVM classifier by replacing it with k-nearest neighbors (k-NN) classification. Prakash *et al.* [10] used k-means for segmenting a leaf image to find infected areas and SVM was used for classification purpose. Bhuvana *et al.* [11] used k-means and multi class SVM for the detection of leaf diseases depend on images dataset and the accuracy obtained equal to 98%. Adit *et al.* [12] applied convolutional neural network (CNN) to analyze plant disease and produced reliable platform. Khan *et al.* [13] examined how ML approaches have been evolved to deep learning for detection of plant disease. Geetha *et al.* [14] used four stages for detect the type of plant disease contained pre-processing, k-means (segmentation of leaf), extraction of feature and classification using k-NN. Sankaran *et al.* [15] used k-means, principal component analysis (PCA) and SVM proposed for detect leaf disease. They describe that accuracy (ACC) of k-means in predicting best from SVM and SVM best from PCA. Mariyappan [16] used k-means and SVM for predicting of disease based on data that including healthy and diseased leaf image. Yousuf and Khan [17] suggested technique world in segmentation by k-means, feature selection by random forest (RF), and classification by k-NN for plant disease. The proposed method outperformed SVM, according to the results of the experiments.

Previous articles have found that k-means, SVM and k-NN are widely applied because of their easy to train, ease of interpretation, and thus have been applied in several articles of classification plant disease. Also have found that k-NN are typically the most accurate than SVM. Depend on the literature; it has been discovered that to date, most articles have applied data of images for plants. But although the literature presents several classification algorithms with k-means clustering for controlling plant disease, none of these articles adopt a complete, integrated, and customized framework that uses an effective clustering algorithm or k-medoids algorithm in plant disease to analyze and forecast plant disease, despite that, clustering algorithms are essential to analyze plant disease data.

This article observes that able to use clustering algorithm in predicting plant disease depend on raw dataset. For observing that we take the same data in the article, which presents by Morgan *et al.* [18] applied classification algorithms on soybean data and observe that k-NN performed best from other classification algorithm in predicting soybean disease where indicated accuracy of k-NN equal to 91.83%. While by implementing our proposed model k-NN achieved 100% accuracy. Therefore, through an in-depth research on classification algorithms and available comparison methods, we have shown that the best way to select an algorithm is to try them all on the same data set.

The rest of this articles is organized as: section 2 discusses the research methodology. Section 3 presents results and analysis. Section 4 discusses the results of this article. Section 5 presents a conclusion as well as recommendations for future research.

2. RESEARCH METHOD

This article aims to improve prediction of a plant disease with a k-efficient clustering and classification model based on data set of soybean that used in [18] to achieve accurate and meaningful results to support agriculture decision-makers. This model enables a better understanding of the nature of the soybean disease. Figure 1 offers block diagram for our proposed model. After we load the data the preprocess

step done, after that the clustering algorithms implemented in the cleaning and balance data, then the k-NN implemented on each disease cluster of soybean data, and finally we evaluating the proposed framework.

We use clustering algorithms before classification algorithms of ML. Since we know, the problem in particular of soybean dataset that containing 35 attributes and 307 instances, the data are multi-type and large. These characteristics present enormous challenges for ML in predicting plant disease. So, we found that integrating of clustering and classification it very suitable in raw data of plant disease. Thus, in this article, we are improving the prediction of plant disease by using an efficient clustering (integrating between k-means and k-medoids) that addresses the key requirements of interoperability, scalability, context discovery, and reliability.

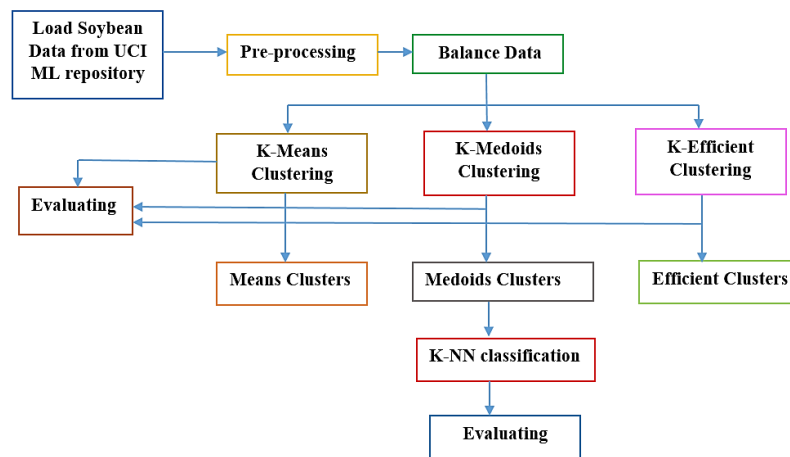


Figure 1. Block diagram of proposed AI model

2.1. Description and preprocessing of soybean data

The soybean dataset, which was downloaded from the UCI machine learning repository, has 307 observations from soybean plants affected with 18 disease and one class for healthy and 35 categorical attributes [19], [20]. These data serve the ML community to investigate their algorithms on a case-by-case basis. According to Google Scholar, the archive has been cited at least 52 times [20]. Soybean data are selected on the basis of their respective large size and its raw data in order to test the clustering algorithms impact on effectiveness, efficiency and scalability of classification algorithms for predicting soybean disease and to show that it is enable for using raw data instead of images for predicting disease of plants.

Soybean dataset include duplicates, outliers, errors (noisy) and some values are missed (incomplete). Data preprocessing is responsible for dealing with these problems. Data preprocessing is a technique of ML and a data mining that converts raw data into a format that ML algorithms can understand [21]. Because of bad data, no meaningful findings are obtained by the ML models or even incorrect replies that can lead to incorrect conclusions.

Before beginning data preprocessing, it's a good idea to figure out what data the ML algorithm need in order to achieve good results. In this article we use the clustering algorithms. Thus, we remove irrelevant observations, duplicates, unnecessary columns, and errors. Also we handle outliers, inconsistent data, and noise [22]. The soybean dataset's attribute values were also numerically coded. A decision was made to run k-NN classification and clustering algorithms on cleaning and balanced soybean dataset.

2.2. K-Means and k-medoids clustering algorithms

The procedure of k-means (unsupervised ML) starts by randomly selecting k (cluster's number) and data points (cluster's center). Then each point is placed in the cluster with the closest center to it. The closest between points and centers calculated based on distance of Euclidean. The centers of the clusters are updated and replaced by the means of the points of the cluster. It is iterated until the clusters have reached stability [23]. It is easy to use, however the selection of the integer k can vary the results [24]. The approach is prone to outliers and noise because the average of the points within the cluster induces early convergence. It also restricts non-numerical data types such as ordinal data and categorical [25].

The procedure of k-medoids is same to that of k-means. But it different from k-means in the step of updatd and replacd medoid by comparing the sum of distances between a point and other points and the sum

of distances between points and medoid. As a result, k-medoids outperform k-means in terms of noise and outliers. Because the use of medoids to represent the centers inhibits early convergence, the findings are relatively successful [26], [27].

We offer k-efficient clustering technique for soybean disease prediction that depends on the merging of k-means and k-medoids. The flowchart in Figure 2 summarizes the algorithm of k-efficient clustering. In theory, the complication is the same as k-medoids. It serves as a bridge between k-means and k-medoids. Since the third instruction executes less iterations, intuitively it is close to the k-means. k-means produce good ones compared to random start centers. As a result, k-efficient clustering can handle large data sets. It is good at least as k-medoids because k-medoids are called at the end of k-efficient clustering. Thus, we can also conclude that it is better than K-means. Because of the simultaneous call of the k-medoids and k-means, there is a drawback of random number k initialization, which is less biased.

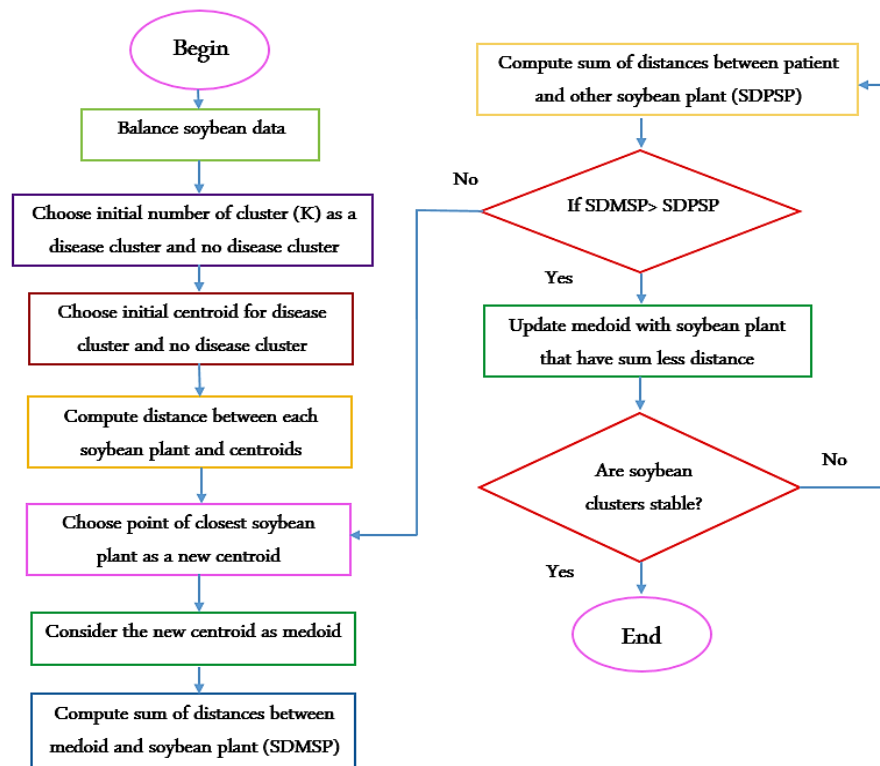


Figure 2. Flowchart of integrating k-means and k-medoids

2.2.1. K-nearest neighbours classification algorithm

k-NN (supervised ML) algorithm is a recognition of pattern that is used in various areas. It is known for its good interpretation, simple procedure, and its low calculation time [28]. The procedure of k-NN depend on the hypothesis that points with same inputs have same outputs. In many data grouping or classification jobs, the distance or similarity between two items plays an important role. For numerical variables in datasets, traditional distances such as Euclidean can be used. The Figure 3 displays a k-NN algorithm flowchart [29]. ACC is more important than any other factor in detecting soybean disease. For the following reasons, the k-NN algorithm may compete with the most accurate models in very accurate predictions [30]:

- When making real-time predictions, k-NN saves the training dataset and solely learns from it. As a result, the k-NN training method is significantly more rapid than the alternatives, such as linear regression, SVM, and so on.
- As no prediction training is necessary for the k-NN, new data may be smoothly inserted that will not influence the algorithm ACC.
- k-NN is a simple algorithm to implement. Only the value of K and the distance function are necessary to implement k-NN [25].

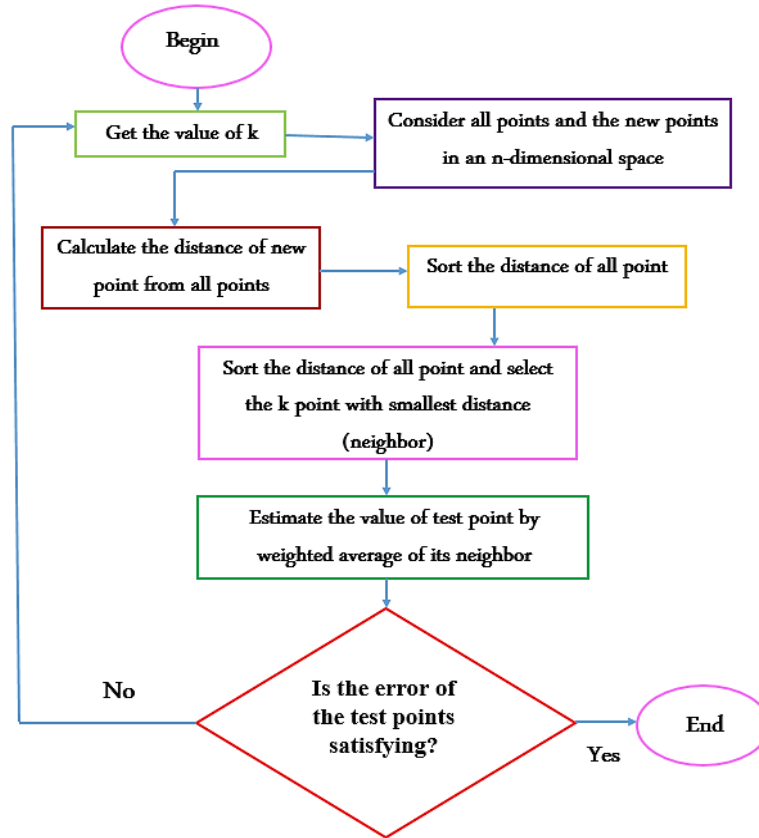


Figure 3. Flowchart of k-NN

2.2.2. Evaluating of clustering and classification algorithms

k-NN is a supervised learning algorithm that is used to solve regression and classification problems while clustering algorithms are unsupervised learning algorithms that are used to solve clustering problems. This is the basic difference between clustering and classification. Thus the evaluating of k-NN is different from evaluating of k-means, k-medoids, and k-efficient clustering [31]. The metrics of evaluation assess the model of classification depend on the ability to accurately forecast the target classes of the unlabeled instances. Forecasting the target classes of the unlabeled instances in a two class issue can be categorized into 4 kinds namely false negative (FN), true positive (TP), false positive (FP), and true negative (TN) as described in Table 1 [32]. The ACC, F-measure, recall, and precision are used for evaluating classification model.

Table 1. Soybean confusion matrix

	No Soybean Disease	Soybean Disease	
No Soybean Disease	TN	FP	Specificity= $TN / (TN + FP)$
Soybean Disease	FN	TP	Recall= $TP / (TP + FN)$
	Negative Predictive Value= $TN / (TN + FN)$	Precision= $TP / (TP + FP)$	Accuracy= $(TP + TN) / (TP + TN + FP + FN)$

The number of correctly classified instances (i.e. TP and FP) splitting by the total number of examples yields ACC. The fraction of relevant instances to the retrieved instances is known as precision. Divide the number of TP classes by the sum of the number of TP classifications and the number of FN classifications to get the recall. Finally, the F-measure is computed by multiplying the recall and precision, dividing this value by the sum of the recall and precision, and finally multiplying this number by two. To demonstrate the effectiveness of clustering algorithms, we used the normal mutual information (NMI) for measuring the degree of closeness of the one that should be in reality to the result of the classification. When the NMI is near to 1, the results are significant. In addition, we employed the inertia to test the algorithm’s effectiveness. Two types of inertia were examined [33], [34]: The inertia of intra class, it is a value to

minimize because it represents the distance between a cluster elements, centroid, and the inertia of interclass, which calculates the distance between the clusters and centers. A high value expresses an interesting result, hence it is a value to maximize. The equations of calculated these metrics have been shown in [34].

3. RESULTS AND ANALYSIS

The classification and clustering algorithms were tested on cleaning soybean dataset in seven experiment, first experiment implementing k-NN without any clustering algorithm, second experiment implement k-means clustering algorithm, third experiment implement k-medoids clustering algorithm, fourth experiment implement k-efficient clustering, fifth experiment k-NN with K-means, sixth experiment k-NN with k-medoids, and finally k-NN with k-efficient clustering. The measures previously determined were reported for algorithm of classification and for clustering algorithms have been used for evaluating the proposed model. Before implementing classification methods to a dataset, a classifier’s ACC can be increased by using clustering techniques. All the experiments were running in Java and on a machine with 2.60 GHz CPU, Core (TM) i5-3230M of 4.00 GB and a RAM processor Intel (R) under Windows 10 64-bit. We put the algorithms to the test with soybean data from the UCI machine learning repository.

3.1. Clustering algorithms

The results of inertia and NMI experiments for the three clustering algorithms shown in Figure 4 and Table 2. K-medoid intra classes are lower than k-means intra-class, and because the distance between positive (each time, one disease type is assigned to the positive class, while the others are assigned to the negative class) and negative clusters in k-means clustering is considerable while the distance between plants and centroid is minimal, the inter-class is larger than k-means. The k-efficient clustering has the shortest distance between positive and negative clusters (intra-class inertia) and the biggest inter-class inertia. In other words, k-efficient clustering is the best strategy for distributing soybean data across disease clusters when compared to other methods. Then we conclude that the inertia of other techniques is exceeded by k-efficient clustering. We refer that k-efficient clustering performs better than k-medoids and k-means in NMI metric. Also, we conclude that k-efficient clustering is practically as well as k-medoids even if k-efficient clustering results in better inertia. Finally, we note that k-efficient clustering takes longer to execute than k-means but less time than k-medoids for forecasting soybean disease. Because it is faster than k-medoids, it fits our scenario. K-efficient clustering is effective in obtaining the best of both techniques for the benchmark: k-medoids precision and k-means speed.

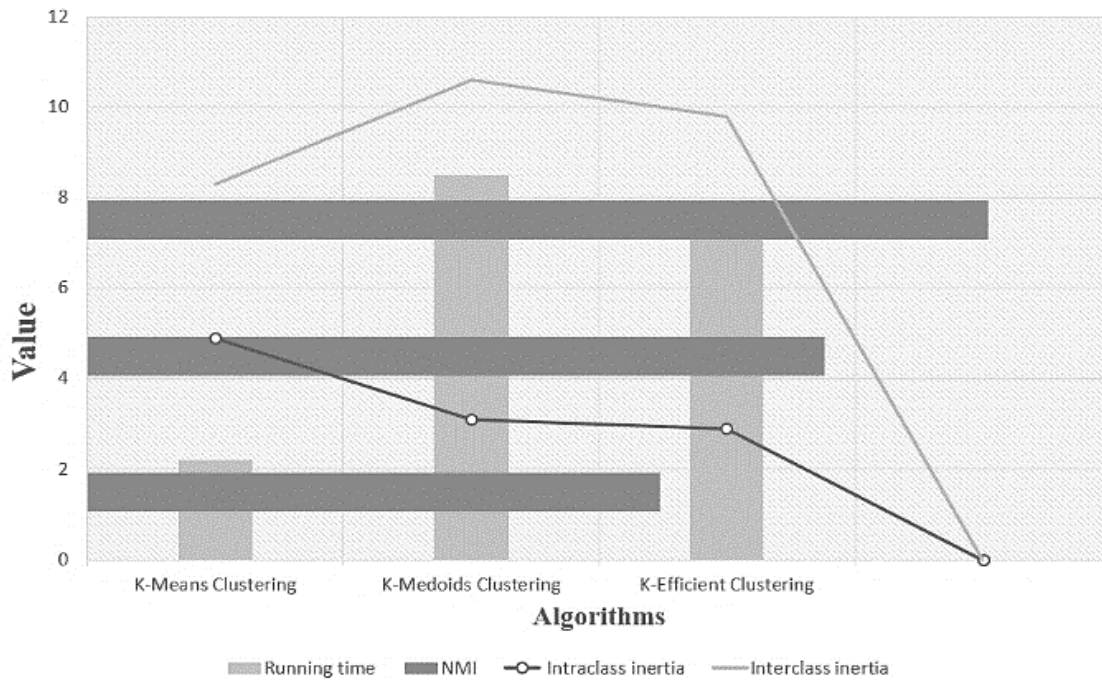


Figure 4. Results of clustering evaluating

Table 2. Results of clustering evaluating

Algorithm	Intra-class inertia	Inter-class inertia	NMI	Running time
K-Means Clustering	4.9	3.4	1.4	2.2
K-Medoids Clustering	3.11	7.5	1.8	8.5
K-Efficient Clustering	2.9	6.9	2.2	7.2

3.2. Integrating clustering and classification

In comparing the results that shown in Figure 5 and Table 3 of four experiments, k-NN performed best on the soybean dataset with k-means from without using clustering, k-NN performed best on the data of soybean with k-medoids from k-NN with k-means while k-NN performed at 100% accuracy with k-efficient clustering. It is to be expected that most of the classification algorithms would perform best with k-efficient clustering because it is integrating the benefits from k-means and k-medoids, despite the data of soybean that was tested has 19 disease classes. When using a k-NN classifier with clustering, the ACC and true positive rate are greater than when using a k-NN classifier alone to predict soybean disease. We would like the precision number to be 100 and the false positive rate to be zero in an ideal scenario. Integration of k-NN with clustering techniques has a greater precision value than simple classification with a k-NN classifier. Considering results presented in Figure 5, F-measure of k-efficient clustering is best from other algorithms.

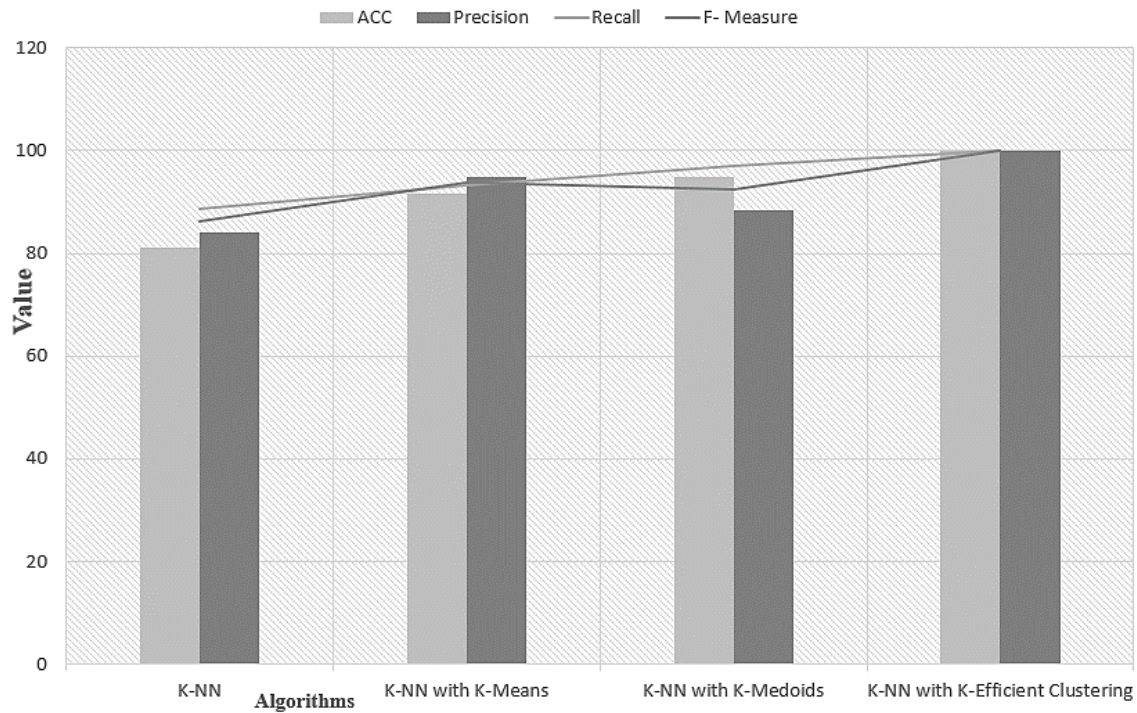


Figure 5. Results of integrating evaluating

Table 3. Results of integrating evaluating

Algorithm	ACC	Precision	Recall	F- Measure
k-NN	81.2	84	88.7	86.2
k-NN with K-Means	91.5	94.7	93.3	93.8
k-NN with K-Medoids	94.7	88.4	97	92.5
k-NN with K-Efficient Clustering	100	100	100	100

4. DISSCUSSION

Integration of k-NN with clustering techniques has a greater precision value than simple classification with a k-NN classifier. To aid in the early diagnosis of plant diseases, rapid and accurate models are required. For this article, we applied a data that included 18 different types of soybean diseases as well as a healthy class. The data was then split into train and test set for evaluating k-NN classification algorithm without clustering algorithms and with algorithms of clustering were all trained applying 10-fold

(cross validation). The rows within the data are randomly restructured and divide into 10 folds with equal size. With each iteration of the training k-NN, one-fold is applied as the test data and 9 folds are applied as the training data. This technique is repeated ten times more until each fold is applied as the test data. The resultant classification model is an average of the training process's 10 iterations. The study worked with seven experiments including k-means, k-medoids, k-efficient clustering, k-NN, k-means with k-NN, k-NN with k-medoids, and k-NN with k-efficient clustering. The evaluating of classification algorithms was different from evaluating the clustering algorithms, whereas k-NN algorithm was subjected to an evaluation depend on the measures: F-measure, precision, recall, and ACC. While evaluating clustering algorithms was based on NMI, intra-class, and inter-class. The results show that k-NN with k-efficient clustering had the best performance while k-NN without clustering, posted the least performance. We expect this new technology will lead to huge new discoveries and be a highly valuable addition to the sector of agriculture.

As shown in the results for the k-NN classification algorithm with k-efficient clustering tested on the data of soybean in Figure 5, the values of recall and precision are equal to 100, this refers that the number of TP classifications are much larger than the number of FN and FP classifications. If the k-NN algorithm were being implemented without clustering, our values of recall and precision values would be much lower.

The results of clustering obtained in this article are similar to those resulted in another article done by Drias *et al.* [34] on breast cancer data set and COIL-100 images dataset. These are used to identify the performance of integrating k-means and k-medoids clustering. The trained k-NN classifier achieved an accuracy of 91.83% in the study implemented by Morgan *et al.*, which compared to our 100% ACC achieved by our proposed model. We compared the result of k-NN our proposed model by Maria study because they used the same dataset. However, the article makes an important contribution by building the possibility of using raw data depend on clustering and classification to predict plant disease. As refers in the section of introduction, which discusses the literature review, k-means with k-NN are almost applied articles of agricultural. We should think about this as k-NN was a high-performing algorithm in the soybean dataset [18]. One of the most significant challenges encountered during the preparation of this study was the lack of data included that the clustering algorithm not implemented truly on data that has been not preprocessed, so the preprocess step was very important and must be implemented for unbalance data. In clustering algorithms, data that are unstable require additional techniques of data preparation as to not skew the findings of the algorithms of classification. It would be interesting to test the proposed model with a different set of raw data, or to see if the models could categorize images of plant diseases as soon as they appeared. Farmers could benefit from using the suggested model, which combines k-NN with k-efficient clustering, to diagnose soybean and other plant diseases.

5. CONCLUSIONS

In this article, we implement seven experiments for testing k-means, k-medoids, k-efficient clustering, k-NN, and k-NN classifier with the three clustering algorithms to forecast presence of disease in classify and predict disease in a data of soybean. In the soybean dataset, we have shown that k-NN with k-efficient clustering are the best classifiers in terms of ACC, but k-NN without clustering have less performance from other algorithms. The goal of these tests was to develop clustering and classification methods that could be applied to plant datasets that contained real measurements rather than images. The results of this study can be replicated using similar plant datasets, and they can also be used to train other classification algorithms with clustering for forecasting classification of disease in animal or human data with raw metrics. This article's work decreases the monitoring effort for large-scale agriculture, and the agricultural scale can profit greatly. This work produces a high-quality product while minimizing the impact on plant productivity and economic profit in the resource progress. Finally, the proposed AI model for improving allowing agriculture to progress.




REFERENCES

- [1] A. M. Abdu, M. M. M. Mokji, and U. U. Sheikh, "Machine learning for plant disease detection: an investigative comparison between support vector machine and deep learning," *IAES Int. J. Artif. Intell.*, vol. 9, no. 4, pp. 670–683, Dec. 2020, doi: 10.11591/ijai.v9.i4.pp670-683.
- [2] S. B. Jadhav, "Convolutional neural networks for leaf image-based plant disease classification," *IAES Int. J. Artif. Intell.*, vol. 8, no. 4, pp. 328–341, Dec. 2019, doi: 10.11591/ijai.v8.i4.pp328-341.
- [3] M. A. I. Aquil and W. H. W. Ishak, "Evaluation of scratch and pre-trained convolutional neural networks for the classification of Tomato plant diseases," *IAES Int. J. Artif. Intell.*, vol. 10, no. 2, pp. 467–475, Jun. 2021, doi: 10.11591/ijai.v10.i2.pp467-475.
- [4] T. U. Rehman, M. S. Mahmud, Y. K. Chang, J. Jin, and J. Shin, "Current and future applications of statistical machine learning algorithms for agricultural machine vision systems," *Comput. Electron. Agric.*, vol. 156, pp. 585–605, Jan. 2019, doi: 10.1016/j.compag.2018.12.006.




- [5] G. Prem, M. Hema, L. Basava, and A. Mathur, "Plant disease prediction using machine learning algorithms," *Int. J. Comput. Appl.*, vol. 182, no. 25, pp. 1–7, Nov. 2018, doi: 10.5120/ijca2018918049.
- [6] E. Fujita, Y. Kawasaki, H. Uga, S. Kagiwada, and H. Iyatomi, "Basic investigation on a robust and practical plant diagnostic system," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2016, pp. 989–992., doi: 10.1109/ICMLA.2016.0178.
- [7] E. Khalili, S. Kouchaki, S. Ramazi, and F. Ghanati, "Machine learning techniques for soybean charcoal rot disease prediction," *Front. Plant Sci.*, vol. 11, Dec. 2020, doi: 10.3389/fpls.2020.590529.
- [8] M. G. Roth *et al.*, "Integrated management of important soybean pathogens of the United States in changing climate," *J. Integr. Pest Manag.*, vol. 11, no. 1, Jan. 2020, doi: 10.1093/jipm/pmaa013.
- [9] G. Kaushal and R. Bala, "GLCM and KNN based algorithm for plant disease detection," *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.*, vol. 6, no. 7, pp. 5845–5852, 2017, doi: 10.15662/IJAREEIE.2017.0607036.
- [10] R. M. Prakash, G. P. Saraswathy, G. Ramalakshmi, K. H. Mangaleswari, and T. Kaviya, "Detection of leaf diseases and classification using digital image processing," in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Mar. 2017, pp. 1–4., doi: 10.1109/ICIIECS.2017.8275915.
- [11] S. Bhuvana, B. K. Bharati, P. Kousiga, and S. R. Selvi, "Leaf disease detection using clustering optimization and multi-class classifier," *WSEAS Trans. Comput. Arch.*, vol. 17, 2018
- [12] V. V. Adit, C. V. Rubesh, S. S. Bharathi, G. Santhiya, and R. Anuradha, "A Comparison of Deep Learning Algorithms for Plant Disease Classification," *Adv. Cybern. Cogn. Mach. Learn. Commun. Technol. Lect. Notes Electr. Eng.*, vol. 643, pp. 153–161, 2020
- [13] R. U. Khan, K. Khan, W. Albattah, and A. M. Qamar, "Image-based detection of plant diseases: from classical machine learning to deep learning journey," *Wirel. Commun. Mob. Comput.*, vol. 2021, pp. 1–13, Jun. 2021, doi: 10.1155/2021/5541859.
- [14] G. Geetha, S. Samundeswari, G. Saranya, K. Meenakshi, and M. Nithya, "Plant leaf disease classification and detection system using machine learning," *J. Phys. Conf. Ser.*, vol. 1712, no. 1, p. 12012, Dec. 2020, doi: 10.1088/1742-6596/1712/1/012012.
- [15] K. S. Sankaran, N. Vasudevan, and V. Nagarajan, "Plant disease detection and recognition using k means clustering," in *2020 International Conference on Communication and Signal Processing (ICCSP)*, Jul. 2020, pp. 1406–1409., doi: 10.1109/ICCSP48568.2020.9182095.
- [16] B. Mariyappan, "Crop leaves disease identification using k-means clustering algorithm and support vector machine," 2020
- [17] A. Yousuf and U. Khan, "Ensemble classifier for plant disease detection," *Int. J. Comput. Sci. Mob. Comput.*, vol. 10, no. 1, pp. 14–22, Jan. 2021, doi: 10.47760/ijcsmc.2021.v10i01.003.
- [18] M. Morgan, C. Blank, and R. Seetan, "Plant disease prediction using classification algorithms," *IAES Int. J. Artif. Intell.*, vol. 10, no. 1, pp. 257–264, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp257-264.
- [19] D. Aha, "UCI machine learning repository." 1987.
- [20] R. S. Michalski and R. L. Chilausky, "Learning by being told and learning from examples: an experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis," *Int. J. Policy Anal. Inf. Syst.*, vol. 4, no. 2, 1980
- [21] M. A. Khan *et al.*, "An optimized method for segmentation and classification of apple diseases based on strong correlation and genetic algorithm based feature selection," *IEEE Access*, vol. 7, pp. 46261–46277, 2019, doi: 10.1109/ACCESS.2019.2908040.
- [22] J. G. A. Barbedo, "Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification," *Comput. Electron. Agric.*, vol. 153, pp. 46–53, Oct. 2018, doi: 10.1016/j.compag.2018.08.013.
- [23] S. H. Toman, M. H. Abed, and Z. H. Toman, "Cluster-based information retrieval by using (K-means)- hierarchical parallel genetic algorithms approach," *TELKOMNIKA (Telecommunication Comput. Electron. Control.)*, vol. 19, no. 1, pp. 349–356, Feb. 2021, doi: 10.12928/telkomnika.v19i1.16734.
- [24] P. Govender and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980-2019)," *Atmos. Pollut. Res.*, vol. 11, no. 1, pp. 40–56, Jan. 2020, doi: 10.1016/j.apr.2019.09.009.
- [25] A. Poompaavai and G. Manimannan, "Clustering study of indian states and union territories affected by coronavirus (COVID-19) using k-means algorithm," *Int. J. Data Min. Emerg. Technol.*, vol. 9, no. 2, p. 43, 2019, doi: 10.5958/2249-3220.2019.00006.5.
- [26] S. A. Abbas, A. Aslam, A. U. Rehman, W. A. Abbasi, S. Arif, and S. Z. H. Kazmi, "K-means and k-medoids: cluster analysis on birth data collected in city muzaffarabad, kashmir," *IEEE Access*, vol. 8, pp. 151847–151855, 2020, doi: 10.1109/ACCESS.2020.3014021.
- [27] F. Rahman, I. I. Ridho, M. Muflih, S. Pratama, M. R. Raharjo, and A. P. Windarto, "Application of data mining technique using k-medoids in the case of export of crude petroleum materials to the destination country," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 835, no. 1, p. 12058, Apr. 2020, doi: 10.1088/1757-899X/835/1/012058.
- [28] D. C. Corrales, "Toward detecting crop diseases and pest by supervised learning," *Ing. y Univ.*, vol. 19, no. 1, p. 207, Jul. 2015, doi: 10.11144/Javeriana.iyu19-1.tdcd.
- [29] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, 3rd ed. 2012.
- [30] R. A. Jaleel, I. M. Burhan, and A. M. Jalookh, "A proposed model for prediction of COVID-19 depend on k-nearest neighbors classifier:iraq case study," in *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, Jun. 2021, pp. 1–6., doi: 10.1109/ICECCE52056.2021.9514171.
- [31] A. A. Amer and H. I. Abdalla, "A set theory based similarity measure for text clustering and classification," *J. Big Data*, vol. 7, no. 1, p. 74, Dec. 2020, doi: 10.1186/s40537-020-00344-3.
- [32] R. Ramya, P. Kumar, D. Mugilan, and M. Babykala, "A review of different classification techniques in machine learning using weka for plant disease detection," *Int. Res. J. Eng. Technol.*, vol. 5, no. 5, pp. 3818–3823, 2018
- [33] H. Drias, N. F. Cherif, and A. Kechid, "K-MM: a hybrid clustering algorithm based on k-means and k-medoids," in *Advances in Intelligent Systems and Computing*, Springer International Publishing, 2016, pp. 37–48., doi: 10.1007/978-3-319-27400-3_4.
- [34] H. Drias, A. Kechid, and N. Fodil-Cherif, "A hybrid clustering algorithm and web information foraging," *Int. J. Hybrid Intell. Syst.*, vol. 13, no. 3–4, pp. 137–149, Feb. 2017, doi: 10.3233/HIS-160231.

BIOGRAPHIES OF AUTHORS






Asraa safaa ahmed    holds a master's degree Computer Science from the University of Diyala in 2019. She also received his B.Sc. (Computer Science) from University of Diyala 2009-2010, My research includes machine learning, data mining, artificial intelligence, brain signal classification, and electromagnetic signal classification. She holds the scientific title of Assistant Lecturer at Diyala University in 2019. Published 6 research papers in international fields and scientific conferences from 2019 to 2021. She can be contacted at email: asraasafaa@uodiyala.edu.iq.






Zainab Kadhm Obeas    she holds a master's degree Computer Science from the University of Diyala in 2019. She also received his B.Sc. (Computer Science) from University of Babylonian 2007, My research includes machine learning, data mining, artificial intelligence, brain signal classification, and electromagnetic signal classification. She holds the scientific title of Assistant Lecturer at Al-Qasim Green University in 2019. Published 4 research papers in international fields and scientific conferences from 2019 to 2021. She can be contacted at email: zainabkadhm1@gmail.com.



Batool Abd Alhade    she holds a master's degree Computer Science from the University of Diyala in 2019. She also received his B.Sc. (Computer Science) from University of Babylon in 2008, My research includes machine learning, data mining, artificial intelligence, brain signal classification, and electromagnetic signal classification. She holds the scientific title of Assistant Lecturer at Al-Qasim Green University in 2019. Published 4 research papers in international fields and scientific conferences from 2019 to 2021. She can be contacted at email: batool@uoqasim.edu.iq.



Refed Adnan Jaleel    received the B.Sc. & M.Sc. degree in Information and Communications Engineering in 2014 and 2020, respectively, from Baghdad University-Al-Khwarizmi-Engineering-College and Al-Nahrain University-Information Engineering College, Baghdad, Iraq. Her research interests are in Internet of Things, Software Defined Networks, Security, Wireless Sensor Networks, Information Systems, Meta-heuristic Algorithms, Artificial Intelligence, Machine Learning, Data Mining, Data Warehouse, Recommender Systems, Image Processing, Cloud Computing, Fuzzy Logic Techniques, and Database Management Systems. She worked as an Editor & Reviewer and she has published many articles in international journals and conference proceedings. She also has many certificates of participation in scientific symposia and electronic workshops. She can be contacted at email: iraq_it_2010@yahoo.com.