

Article

Classification of Leaf Disease via Deep Neural Network combined with Clustering Algorithm

Emma Qumsiyeh^{1*}, Muath Sabha²

1 Computer Engineering Department, Faculty of Engineering and Information Technology, Palestine Ahliya University (PAU), Bethlehem, Palestine

2 Faculty of Engineering and Information Technology, Arab American University, Palestine

* Correspondence: e.qumsiyeh@paluniv.edu.ps

Abstract: The agricultural sector in Palestine has a significant role in its economy. However, the production of this sector is affected by different kinds of plant diseases, specifically leaf diseases. Automatic agricultural leaf disease detection is essential for the early diagnosis and controls the overall health of fields. Image segmentation techniques, clustering, and deep learning are often used to detect diseased leaves. This study proposes a novel hybrid approach based on image classification. The hybrid approach combines the k-means clustering algorithm with Convolutional Neural Network (CNN), where k-means is used to detect the leaf's infected area, then CNN is used for specifying the disease. We used the PlantVillage dataset for experimental verification as it contains several crops with different kinds of challenging diseases. We also examined the selection of optimal k-value using the Silhouette coefficient, Elbow method, and Kneedle Algorithm. The Silhouette technique was analyzed using three distance metrics; Euclidean, Manhattan, and Cosine. Its scores for the three-distance metrics were low, near-zero, and failed to produce the optimum k value. Besides, the Elbow method was complicated to use in image segmentation in terms of executing and visualizing the k value in its graph plot. Based on verification results, the Kneedle Algorithm produced better results in the consistency of choosing the optimal k value and showed superiority over other approaches. Therefore, the processed images were segmented with the k-means clustering algorithm with a Kneedle algorithm-based k value. Finally, a Convolutional Neural Network (CNN) is trained to classify the type of disease based on analyzing and testing leaf images. The hybrid model achieved high accuracy of 93.79% in disease identification, confirming the proposed model's robustness.

Keywords: K-means clustering algorithm; Elbow method; Silhouette technique; Kneedle Algorithm; Image Segmentation; Conventional Neural Network.

1. Introduction

Agriculture is the backbone of the economy in the State of Palestine. Farming, in specific, is one of the largest sectors of agriculture. The production of goods for agriculture not only supports Palestinian living but also empowers the country's economy. In comparison to Jordan and Israeli production, and despite their identical soil and climate, Palestinian agricultural yields are considered very low (Mariele, 2020). Agricultural improvements are critical, and the Palestinian government should put effort into developing this sector to create more livable conditions for the Palestinian community.

Palestine is one of the leading producers of citrus, cherry, grape, and tomato. However, such plants suffer from diseases such as black rot, rust, powdery mildew, gray leaf spot, bacterial spot, and mosaic virus (Sharma et al., 2020). These diseases, if not treated, may spread from one plant to the entire crop. With the vast fields and the huge number of crops, the manual detection of such diseases is challenging. Identifying and classifying leaf diseases is becoming hard for human eyes, and the process of automating it has become indispensable.

Many plant diseases exist; however, we are interested in leaf diseases. Leaves are the core of plants; they compromise the whole plant life cycle to danger when infected. Heterogeneous diseases may affect the leaf leading to infecting the entire plant. The lack of knowledge of such diseases accredits farmers for using pesticides on diseased plants that may affect human lives. Hence, identifying leaf diseases is essential for disease management. Here come Artificial intelligence, machine learning, and image processing techniques for such help. Automating leaf disease detection using image processing techniques is a promising solution (Hassan et al., 2021; Prakash et al. 2017). The fast and accurate detection and classification of leaf disease are needed to strengthen the field of agriculture (Devaraj et al., 2019). As reviewed in S & Raghavendr (2019), the potential of image processing and machine learning techniques in detecting leaf diseases is high, which brings to a close the importance of carrying out this research for the next level of proficiency.

A need to detect leaf diseases at the initial stage is vital. In this research, disease detection is done using four main stages: image acquisition, image preprocessing, image segmentation, and classification using deep learning models. An extensive collection of healthy and unhealthy leaf images is used via the PlantVillage dataset. Some preprocessing is done on these images to remove noise and resize them. Segmenting the exact diseased spots in the leaf is essential in the identification process. The segmentation is done using the k means clustering algorithm. To choose the optimal value of k for each crop type, the Kneedle Algorithm is used. This step improves the accuracy of detecting unhealthy leaves and diagnosing their disease. In the final step, the output of the k-means clustering algorithm serves as the input to the artificial neural network algorithm, Convolutional Neural networks (CNN), which is used to train and test the dataset based on accuracy in predicting the diseased leaf.

2. Literature Review

With technological advances, many techniques have been developed in image processing techniques. Image segmentation is considered one of the most critical stages in computer vision. This step helps understand the image entirely, reduces the complexity, and facilitates the work of other high-level processing tasks. Image segmentation considers the digital image as a set of pixels, where we classify similar pixels into regions. Pixels within the same region tend to be similar and vary from other pixels in different regions (S. Jain et al., 2015). We perform this act to remove noise, isolate backgrounds and identify objects within the same image (Anand et al., 2016; Han 2017).

Various segmentation techniques are used, such as edge detection, thresholding, and data-clustering, as proposed in Yang et al. (2020). Clustering is a preferable technique used in image segmentation, and in specific K-means clustering algorithm is one of the most popular unsupervised algorithms in data mining literature. In clustering, we divide several objects into groups. Objects in the same cluster should be as close to each other as possible and different from other objects in other clusters. The number of k points is randomly selected. Samples are then divided into groups based on the nearest center points. The center point is defined as the center of the cluster until it is unchanged and is the center of all the samples within the same cluster (Li et al., 2017).

Since the k-means clustering algorithm is simple and performs fast results, it is a popular image segmentation method. It has been used extensively in agriculture to detect, diagnose, and prognosis different plant leaf diseases. Much research has been done in the literature in this area. Khalid et al. (2020) used the k-means clustering algorithm in Food image segmentation. The authors presented a novel approach called Hk-means (Homogeneity test of k-means) to prevent the wrong choice of k value when performing k-means clustering, leading to local minima. They first consider a random number of k, then the variance value is calculated. The cluster is homogenous if this value is lower than the homogeneity test. Otherwise, the cluster is nonhomogeneous and consists of more than one

food. Their new algorithm gave an accuracy of 96% in predicting the number of food items on a plate.

Febrinanto et al. (2019) used the K- Means method to segment the disease in an image of a citrus leaf. After segmentation, they used the K-Nearest Neighbor (K-NN) algorithm to distinguish between diseases. Silhouette Coefficient was used to check the quality and strength of a cluster. They studied the optimal number of clusters in two stages; leaf segmentation and disease segmentation. In the first stage, they separated the leaf from the background. The second is the most critical stage, where the diseased leaf area is separated from the healthy part.

Anand et al. (2016) tested brinjal leaf diseases using image processing. K-means clustering algorithm was used for segmentation and Neural-network for classification. The k-means clustering algorithm was used to segment the leaves into clusters. Then the number of diseased clusters was computed to distinguish the severity of the disease. The authors converted the diseased leaf or cluster from RGB to HSV. Then feature extraction was performed using the color Co-occurrence Method (CCM method) to extract the texture, angular moment, the intensity of covariance, and entropy features. Authors have not declared their models' accuracy results.

In this study, we have used the k-means clustering algorithm to segment the leaves images based on the previous studies. Several preprocessing steps are performed, such as image resizing and color space transformation from RGB into HSV. The k-means clustering algorithm was used to segment the leaves into clusters. The optimal k-value was determined using the Kneedle Algorithm. The segmentation helped in identifying the disease parts in the leaf. After segmentation, we used the convolutional neural network (CNN) algorithm to train and test our model to distinguish between diseases.

3. Materials and Methods

3.1. Dataset

The dataset used in our study is the PlantVillage benchmark dataset (Hughes & Salathe, 2016). It is enormous as it contains 54,305 leaf images divided into 38 folders, including healthy and diseased leaves. The images are divided into 12 healthy crops and 26 crop diseases. The diseases are black rot, rust, bacterial spot, early blight, late blight, leaf scorch, target spot, and mosaic virus. These diseases are categorized into five subgroups: bacterial, viral, fungal, mold, and mite disease. The crops are apple, potato, tomato, grape, strawberry, and corn. The images of this dataset are clear as each image contains one leaf image on a homogeneous background. Fig.1 gives a histogram plot showing the distribution for the 38 crop-disease/ crop-healthy folders and the number of samples in each folder. Random samples of images from the PlantVillage dataset are shown in Fig. 2.

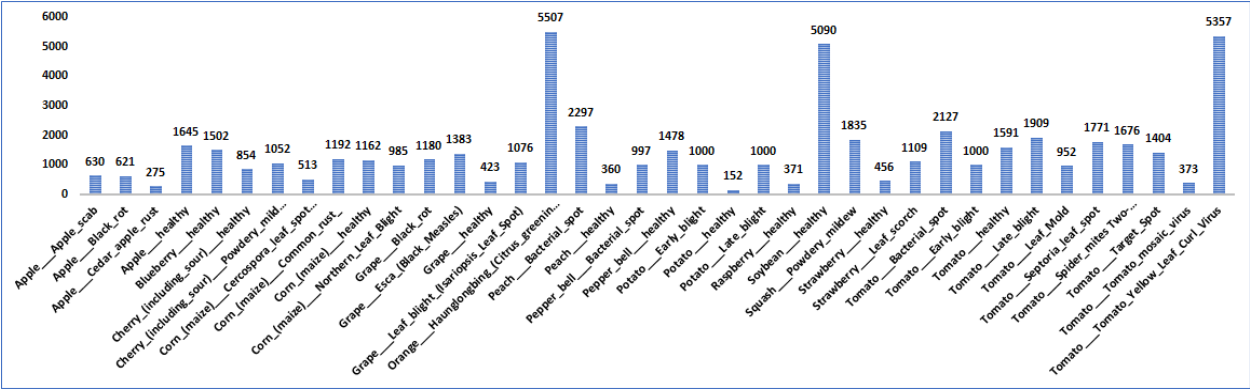


Fig. 1: The PlantVillage dataset Histogram frequency plot, where the X-axis is the crop-disease/ crop-healthy folder name, and the Y-axis is the number of samples.

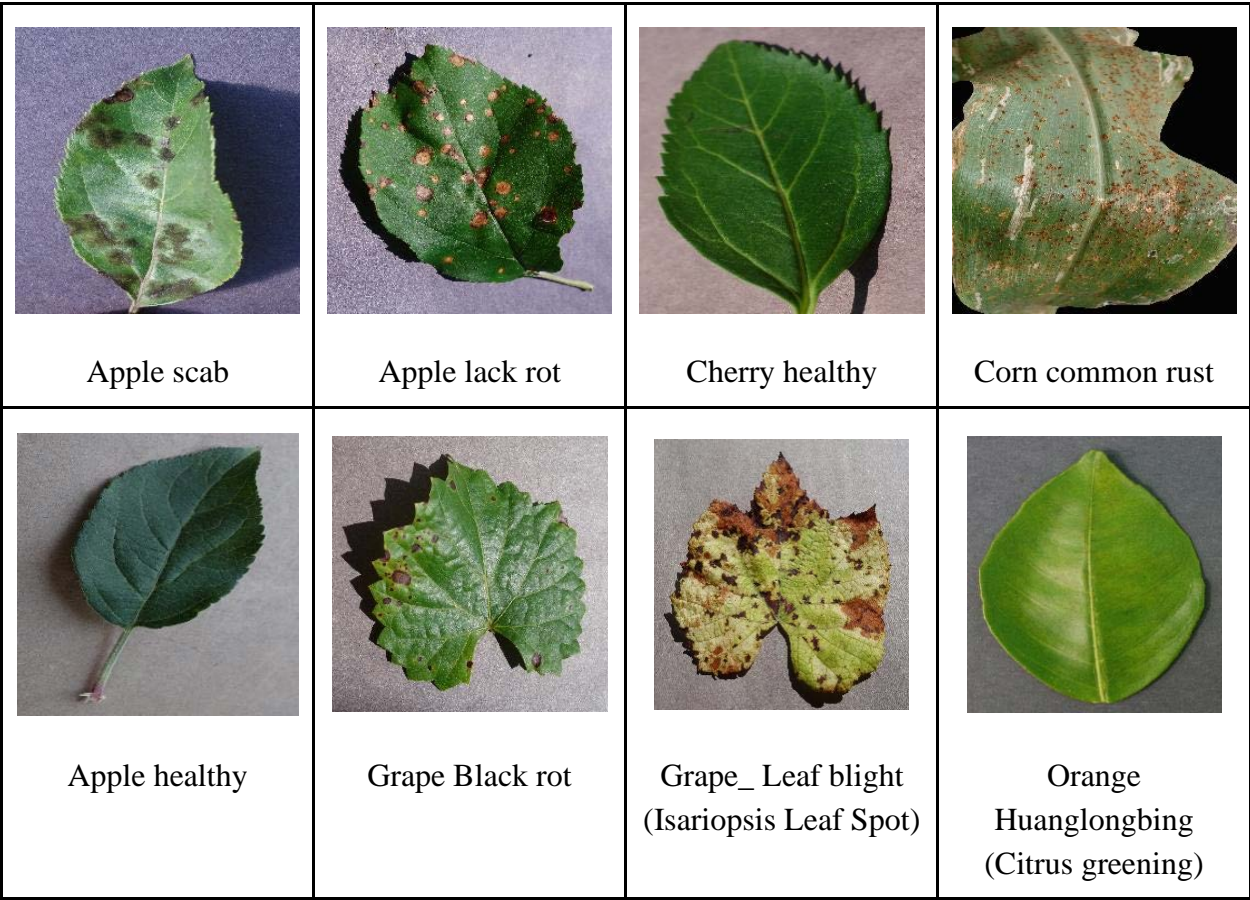
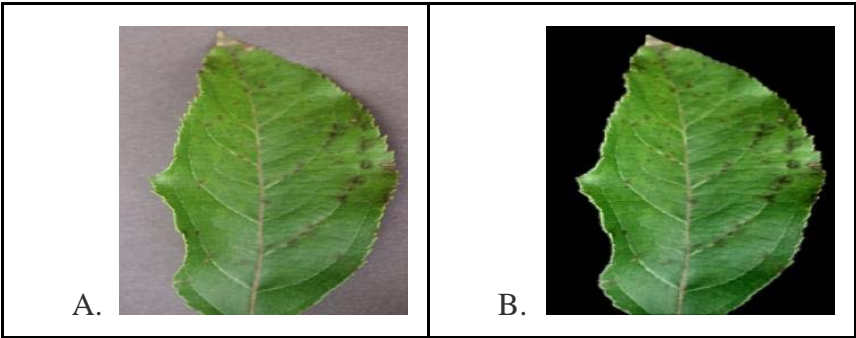


Fig. 2: Random samples of images from the PlantVillage dataset.

The PlantVillage dataset contains three versions; raw images, segmented images, and grayscale images. The segmented ones are RGB images, with the leaf image only extracted from the whole image. This means background removal and color-correcting are performed on images. This process has proven successful in providing clear information, facilitating image analysis, and removing bias from the standardized dataset collection process (Hassan et al., 2021). Fig. 3 represents selected images in RGB and their segmented versions.



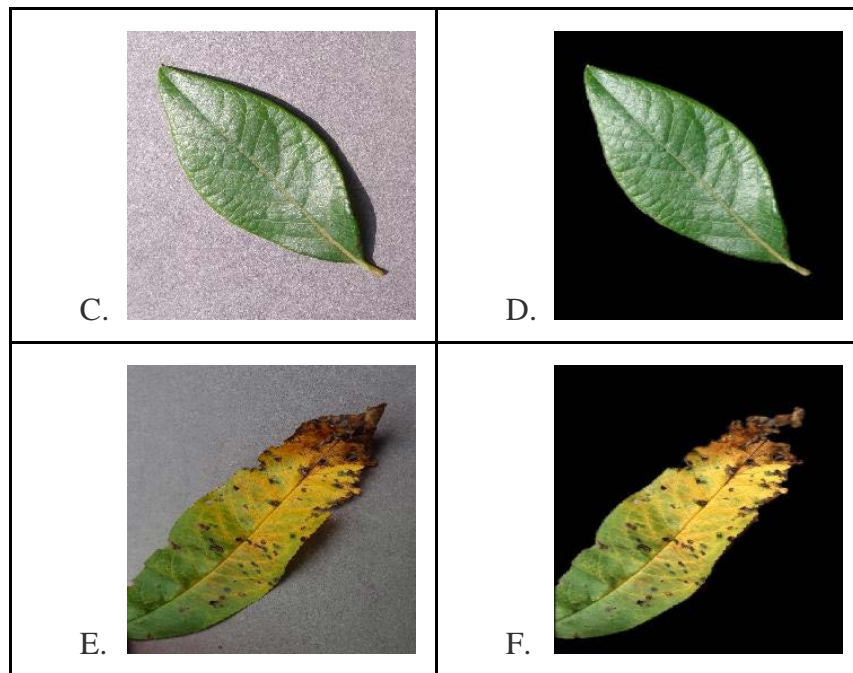


Fig. 3: Sample images from two versions of the PlantVillage dataset used in our experimental configurations. (A) Leaf 1 color Apple scab, (B) Leaf 1 Apple scab segmented, (C) Leaf 2 color Blueberry healthy, (D) Leaf 2 Blueberry healthy segmented (E) Leaf 4 color Peach bacterial spot, (F) Leaf 4 Peach bacterial spot segmented.

As evident in Fig. 3, the segmented version of the images is in good condition. The diseased area in Fig. 3 (F) is reserved in the same state as in the color version (E). No lost information or pixels were observed. These masked segmented images are smoothed to reserve the details needed for leaf disease identification.

3.2. Pre-processing of the dataset

Some preprocessing steps for all the images in the PlantVillage dataset were performed to improve the diagnosis of diseases in plant leaves. Since images are different sizes, we resized them into 256 X 256 resolution. Next, color space conversion is also needed. All images were transformed from RGB color space into Hue Saturation Value (HSV) color space (Vadivel et al., 2005). HSV has proven successful as it is close to the human color perspective. This color space has been used in many studies for leaf disease diagnosis and classification (Arivazhagan et al., 2013; Ramesh & Vydeki, 2020; Rangarajan & Purushothaman, 2020). It also helps identify the leaf image's diseased and non-diseased parts.

3.3. The K-Means Clustering Algorithm

The k-mean clustering algorithm is a simple, iterative, and well-known unsupervised clustering algorithm (Cam & Neyman, 1967). It clusters data into k-clusters. The k should be pre-defined to group the data into distinct non-overlapping clusters. It uses the distance metric to calculate the minimum distance between each data and centroids. Each data belongs only to one cluster, where data in one cluster are similar to each other and different from the other data within other clusters. Fig. 4 illustrates the k-means clustering algorithm scheme.

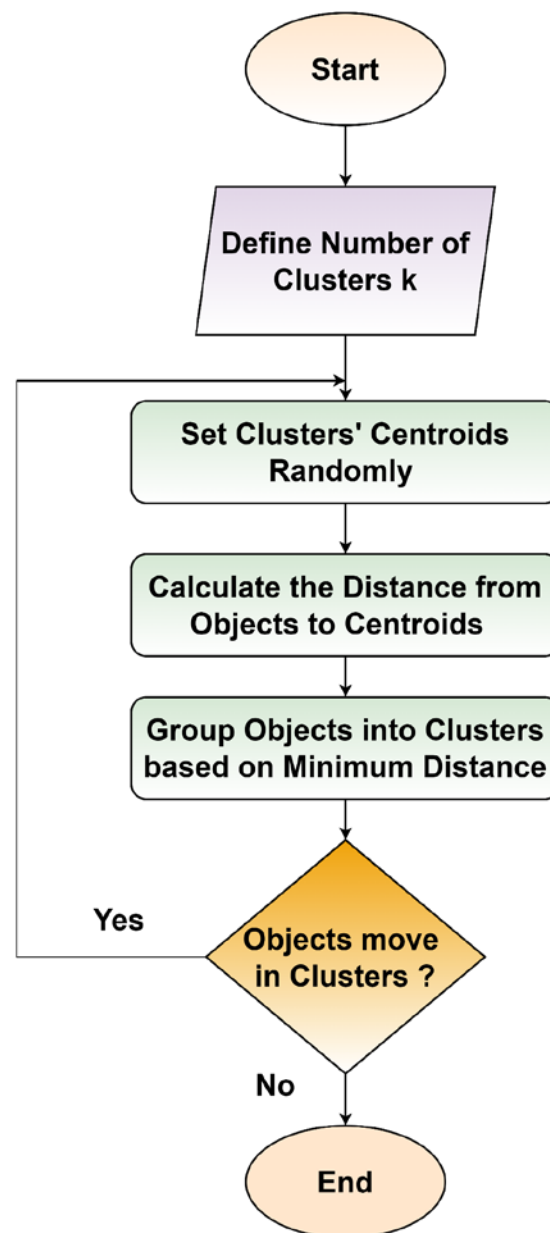


Fig. 4. The K-means Clustering Algorithm workflow

The algorithm starts by randomly determining the number of k clusters. The value of k presents the number of clusters. Then the centroids are placed into these clusters. Creating the clusters uses the distance metric to calculate the distance between each data point and the nearest centroid; then, it places the data into the nearest cluster according to the minimum distance. The centroids keep changing with respect to the data within the cluster it belongs to. The distance is then measured again to the newly updated centroids for each data. The previous steps keep iterating until the centroid value is not changed or updated (Shukla, n.d., p. 14; Sreedhar et al., 2017).

Our proposed method uses the k-means clustering algorithm to detect the diseased tissue in a leaf image. Meaning that segmentation is done using the k-means algorithm. Since it is an unsupervised learning algorithm, it is difficult to determine the k value before processing. The traditional k-means algorithm randomly selects the number of k values from the dataset. Determining the correct value is performed by trial and error, which is inefficient and time-consuming. However, the proper k value significantly impacts the clustering results. If it is determined as a value less than optimal, it produces output that

neglects some critical knowledge in the data. On the other hand, a greater value of k produces overlapping between clusters and unnecessary association (Humaira & Rasyidah, 2020).

Three methods are well known and used to determine the optimal value of k : the Elbow Method (Syakur et al., 2018), the Silhouette Coefficient method (Starczewski & Krzyżak, 2015), and the Kneedle Algorithm (Satopaa et al., 2011). The first two methods, the Elbow and Silhouette, produce a graph that can be used to determine the optimal value of k (Yuan & Yang, 2019). The Elbow has a problem determining the inflection point in its graph as it is sometimes not obvious. The Kneedle algorithm is a computational method that is more straightforward in determining the optimal k value.

Another metric that affects the clustering result is the distance metric in the k -mean algorithm. Distance metrics play a significant role in clustering performance. The traditional k -means algorithm uses the Euclidean metric (Gupta & Chandra, 2021). However, different distance metrics like Cosine, Mikawaski, and Manhattan have shown different outputs in the clustering results. Therefore, determining the right metric results in enhancing the overall clustering accuracy. In the Silhouette Method calculation, we have evaluated three distance metrics, Euclidean, Manhattan, and Cosine. Results varied in terms of the optimal value of k and executing time.

3.4. Optimal Cluster Value Analysis

The optimal cluster value is analyzed in this section. It is crucial to consider each disease type and its specification for disease identification and diagnosis. The PlantVillage dataset is enormous; the leaves' shapes are different from each other. Some leaves vary in size; they vary from small square meters to large ones (Greene et al., 2015). Their shape feature also differs from needle shape to oval shape. At the same time, margins can be symmetrical, teethed in shape, sharp points, or even widely wavy. More important to mention is the vein shape as it can be arcuate shape, parallel, palmate, or rotated (Munisi-ami et al., 2015). Nevertheless, the disease affects the shape of the leaf. Diseases like mosaics display a range of white dots on the leaf area. Other diseases like black rot, which affect the grape, turn some leaf parts into yellowish and brownish (A. Jain et al., 2019).

To consider all of the above changes in leaf features and obtain good results, we studied each type of crop-healthy and crop-diseased plant separately by implementing the Silhouette method, the Elbow method, and the Kneedle algorithm on each folder in the dataset. This means we have analyzed the 38 folders separately. We also considered different distance metrics like Euclidean, Manhattan, and Cosine to find the optimal value of k using the Silhouette method. Further, the Elbow method was applied, and the Kneedle algorithm was used to analyze the Elbow graph plots computationally. The details of our approach are described in the following sections.

3.4.1. The Silhouette Method

Rousseeuw (1987) first proposed the silhouette coefficient. It has two main factors; cohesion and separation. Cohesion represents how similar each point is to its cluster. In comparison, separation calculates the difference between each point in one cluster compared to all other clusters. The silhouette coefficient value is in the range of 1 to -1. The high value close to 1 indicates that the data point firmly belongs to its cluster and is far from other clusters. Besides, if the mean value for all data points within the same cluster is high, then the cluster structure is optimal and appropriate. The value 0 means that clusters are overlapping. While the value -1 indicates that the structure of the cluster is improper, and the value of k is either low or high and not optimal. Equation 1 presents the Silhouette coefficient formula:

$$s(i) = \frac{b(i) - a(i)}{\max \{a(i), b(i)\}} \quad (1)$$

Where:

$a(i)$ is the average distance between the data point i and all the other data points within the same cluster. The smaller the value of $a(i)$, the more the data point i belongs to its cluster.

$b(i)$ is the average distance between the data point i and all the other data points that belong to other clusters, not the one it belongs to. The larger the value of $b(i)$, the less the data point i belongs to this cluster.

$s(i)$ is the Silhouette coefficient for the data point i .

As seen from Equation 1, the silhouette coefficient is calculated for every data point within the cluster. Further, the average silhouette score for every k in the cluster is calculated as the mean of $S(i)$. The model then plots a graph between the mean of the cluster's Silhouette and k . From it, we can obtain the optimal value of k for the cluster.

The Silhouette coefficient depends on calculating the distances between the data point and centroids. Different distance metrics, such as Euclidean, Manhattan, Minkowski, and Cosine, can be used in these terms.

The Euclidean distance is a metric that measures the distance between two data points. It is used in the traditional k-means clustering algorithm (Dattorro, 2005). Euclidean distance could be obtained from the square root of the sum of the square differences between two data points, as shown in Equation 2:

$$Euclidean\ Distance(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2} \quad (2)$$

Where: the n value is the total number of data points in the cluster. x and y are two data points.

The Manhattan distance is also called the city block. It is a metric to measure the distance between two data points. It is the sum of the lengths of the line segments between two data points onto the coordinate axes (Craw, 2010). Its formula is shown in Equation 3 below:

$$Manhattan\ Distance(x, y) = \sum_{i=1}^n |x_i - y_i| \quad (4)$$

Where: the n value is the total number of data points in the cluster, x and y are two data points.

The Cosine distance between two data points is the complement of the cosine similarity. It is one minus the cosine of the angle between the vectors of two data points. It depends on the through orientation measurements between points, and since we are dealing with image segmentation, this metric fits our proposed approach. It has been widely used in image segmentation literature and has shown promising results (Bora & Gupta, 2015; Cui et al., 2021). If two vectors are the same, then the cosine of the angle between these vectors is zero; thus, the distance is zero.

The cosine distance between the vector y and x is illustrated in Equation 4 as follows:

$$Cosine\ Distance(x, y) = 1 - \frac{x \cdot y}{||x|| \cdot ||y||} \quad (5)$$

3.4.2. The Elbow Method

This heuristic method is used to find the optimal value of clusters by plotting a graph. The Elbow plot is a data visualization method that is generated by fitting the k-means

algorithm on a range of different k values and then plotting the SSE for each cluster. It calculates the sum of square errors within the cluster (SSE) for each k value. To determine the optimal value of k , the visualization of the graph is analyzed to find the inflection point. It is selected when the SSE value considerably drops on the curve, forming a slight elbow angle (Syakur et al., 2018). Defining k within the cluster means that we can not add additional information to it, and choosing a higher value of k , makes it challenging to separate the clusters more.

It has been reported that the Elbow method is an ambiguous technique as sometimes it is hard to define the right value of k in the plot. Moreover, it does not provide a measurement metric explicitly showing the optimal Elbow points (Shi et al., 2021). We implemented it on random images for each crop folder, not the whole dataset. Since plotting the optimal k in the Elbow method is sometimes ambiguous, we referred to the Kneedle Algorithm to choose the optimum k computationally from the Elbow graphs produced.

3.4.3. The Kneedle Algorithm

The “Kneedle” algorithm or Knee point detection, first presented by Satopaa et al. (2011), is a generic tool to detect the valuable data points showing the best balance trade-offs, called “knees” when the curves have negative concavity or also called “elbows” when the curves have positive concavity. The kneedle algorithm uses the mathematical curvature measure to detect how much a function differs from a straight line by showing the points of maximum curvature in any data set, where the points of local maxima in a set of points in a continuous curve function differ most from a straight line. Simply, knees occur when a curvature decreases and the curve becomes flatter.

4. Results and Discussion

The main aim of this study is to identify and classify leaf diseases in different crop-leaf images using the k -means clustering algorithm and deep learning models. Since obtaining the optimal k value in such an algorithm is an issue, we used the Silhouette method, the Elbow method, and the Kneedle algorithm to determine the best k value to identify healthy plant tissues from unhealthy ones. The Silhouette method has been performed using different distance metrics to evaluate their effectiveness. The methods’ performances are analyzed and discussed in terms of best clustering results and metrics. This section presents the results obtained.

All implementations are carried out on the PlantVillage dataset presented in Fig. 1. The simulation was performed using Python 3.9.0 using an Intel(R) Core(TM) i7-10700 CPU 2.90GHz device with a 16 GB RAM platform (Windows 11 64 bit).

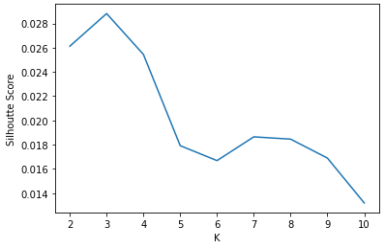
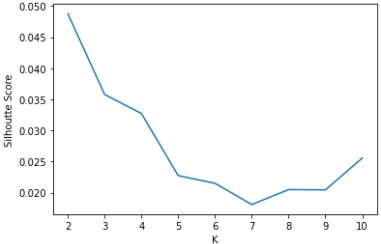
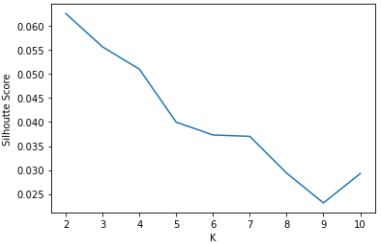
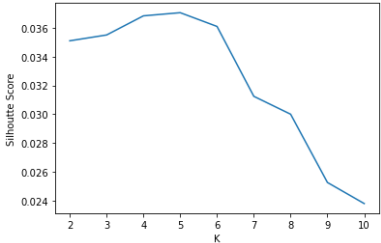
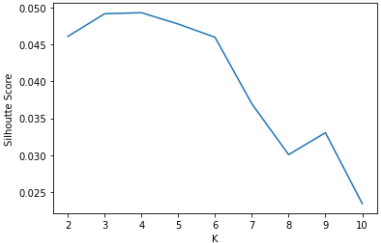
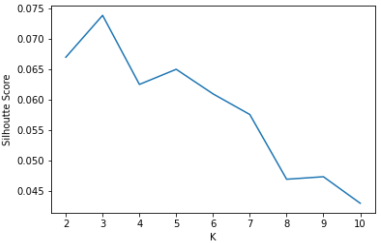
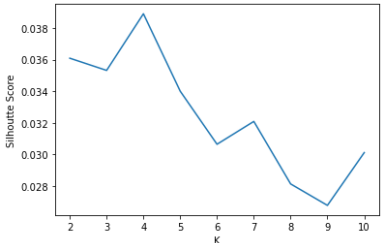
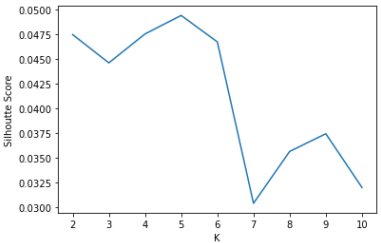
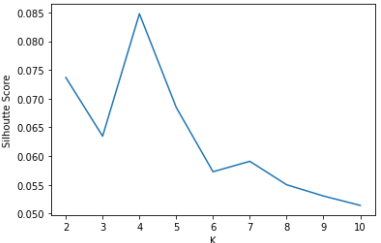
4.1. The Silhouette Method Results

All the results of the Silhouette method are presented in the supplementary files. According to the results and in terms of defining the optimal value of k , the three distance metric results are somewhat similar. Only six results have an un-similar k value out of thirty-eight. For example, for the Apple Scab dataset, the optimal k value by the Silhouette method using the Euclidean distance was 3, whereas the Manhattan and Cosine distance gave the result of 2. Moreover, for the Apple Cedar rust, the optimal k defined by the Silhouette method using the Euclidean distance is 5, whereas the Manhattan and Cosine distance defined it as 4. As evident, the optimal k value result differences for these six datasets are not critical and almost close to each other. This is also clear in Table 1 as it gives the plot curves of the Silhouette method for the Apple Scab, Blueberry healthy, and Tomato mosaic virus folders for the three distance metrics while performing the Silhouette method.

The only difference is that the execution time differs between the three metrics. The Manhattan distance is a 5-fold difference between the Cosine and Euclidean. In addition, the Cosine and Distance performance is close in terms of execution time. We can conclude

that the distance metric has little impact on defining the optimal value of k in the silhouette method. Our finding is also consistent with the results found by (Gupta & Chandra, 2021). Concerning the execution time, the Manhattan distance is computationally heavy and time-consuming since it has the slowest speed with a 5-fold time more than the Cosine distance.

Table 1: The plot curves of the Silhouette method for 3 random folders of the PlantVillage dataset.

	Silhouette method using Euclidean distance	Silhouette method using Manhattan distance	Silhouette method using Cosine distance
Apple scab			
Highest peak/ Score	K value= 3 Score= 0.03	K value= 2 Score= 0.049	K value= 2 Score= 0.07
Blueberr y healthy			
Highest peak/ Score	K value= 5 Score= 0.038	K value=4 Score=0.049	K value= 3 Score 0.073
Tomato mosaic virus			
highest peak/ Score	K value= 4 Score= 0.04	K value= 5 Score= 0.048	K value= 4 Score= 0.085

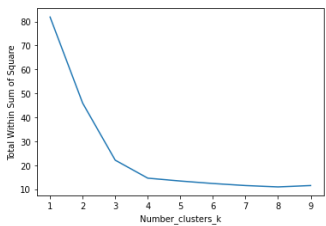
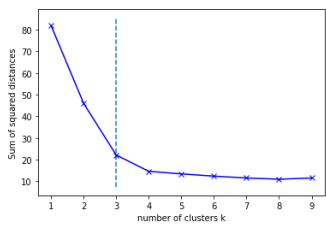
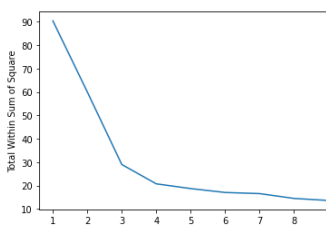
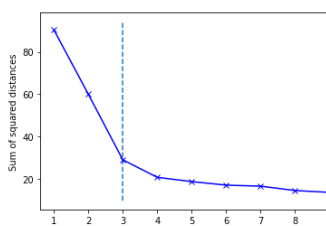
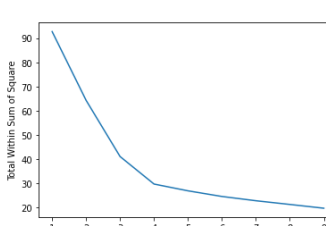
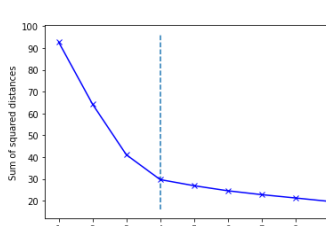
However, as shown in Table 1, the Silhouette score for all the folders using the three distance metrics is zero. A zero value indicates that the clusters overlap, where some samples are very close to the boundary between two neighboring clusters (Nanjundan et al., 2019). This means that the Silhouette method fails in producing the optimal k value, and we can not depend on it to evaluate the optimal k for the 38 folders.

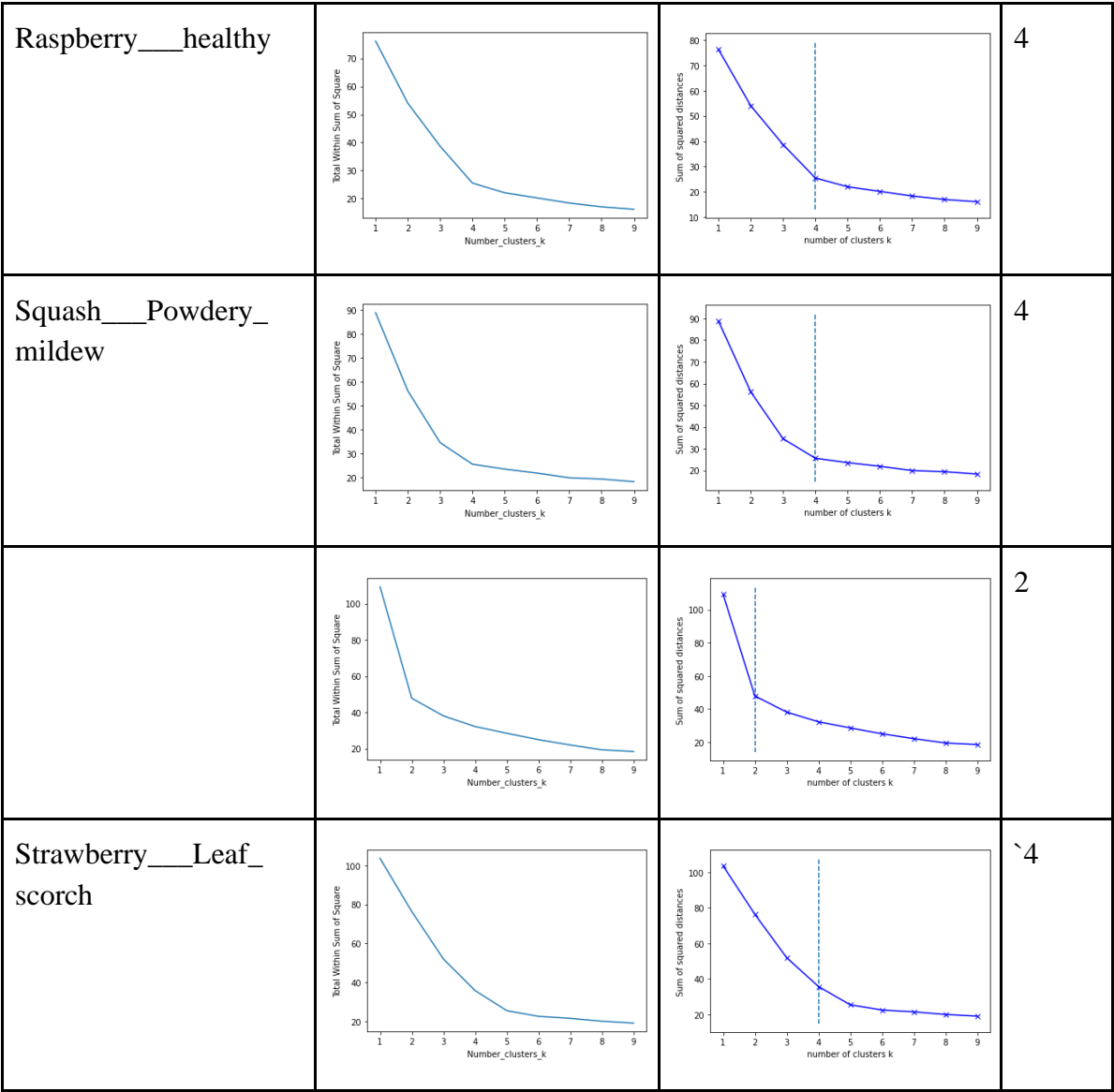
4.2. The Elbow Method Results:

Table 2 presents the curve plots of the Elbow technique and Kneedle algorithm for 6 random crop images in the PlantVillage dataset. Column 1 presents the folder's name. Column 2 shows chosen images. Column 3 presents the plot results of the Elbow method. Column 4 presents the plot from the Kneedle algorithm result. Finally, the optimum k value is identified in column 5.

As seen in Table 2, the Elbow Method algorithm uses the sum of squared errors as a performance metric. It traverses the K value and finds the inflection point. The inadequacy is evident when the inflection point is hard to detect. For example, the Elbow output graph for Cherry_including_sour image in Row 2, Column 3 shows that the highest curvature of the Elbow can be either three or four. Therefore, we referred to the Kneedle Algorithm, and for this specific image, the algorithm's output is presented in column 4. The Kneedle algorithm output graph draws a vertical line cutting the best Elbow at k value =3, where the curvature seems to be stable after it and does not change much. Therefore, the best number of clusters for this image is 3.

Table 2: The Elbow and Kneedle algorithm graphs for 6 random crop-disease/ crop-healthy images along with the k-value specified.

Folder	The Elbow method	The Kneedle Algorithm	K value
Cherry_(including_sour)			3
Grape__Black_rot			3
Potato__Early_bligh			4



It is evident in Table 2 that the optimum k value differs between the different images in the PlantVillage dataset due to the wide variety of diseases and crops. Besides, the Elbow method for different images for a specific crop infected with a disease produces different Elbow plots with different Elbow inflection points. This is evident in Table 2 for *Squash___Powdery_mildew* where the first image has a curvature point with a k value of 4 and the second image with a k value of 2. Therefore, analyzing each folder within the dataset seems to be the best method for our model.

4.3. The Kneedle Algorithm Results:

We have executed the Python package “kneed” (Satopaa et al., 2011) to compute the optimum k value without the need to visualize the Elbow plot. We applied the “kneed” package to every single image within each folder of the PlantVillage dataset. Further, we saved the value of k for each image within the same folder into an array to find the mean k value for each folder. We repeated the previous step for the 38 folders of the PlantVillage dataset. The optimal k value is then calculated using the Kneedle Algorithm to obtain the optimal value of k for each folder. Table 3 presents some results of the Kneedle algorithm. We also added the results of the Silhouette score to compare the two methods. As explicit in Table 3, the mean value of optimal k obtained by the Kneedle Algorithm is higher than those obtained by the Silhouette method. This identifies the zero score in the Silhouette method, which means that clusters are overlapping, meaning the k value should be

higher. The Kneedle algorithm output a higher k value than the Silhouette and succeeded in obtaining a good clustering process. All the detailed results of the Kneedle algorithm are added in the supplementary files.

Table 3: Some results of the optimal k value obtained from the Kneedle Algorithm. The k value obtained from the Silhouette method is also presented for comparison verification.

Folder	kneedle Algorithm		K value by Silhouette coefficient
	k value	mean k value	
Cherry_(including_sour) healthy	2.791569087	3	3
Cherry_(including_sour) Powdery_mildew	3.975285171	4	3
Corn_(maize) Cercospora leaf spot Gray leaf spot	4.040935673	4	2
Corn_(maize) Common rust	4.651006711	5	2
Corn_(maize) healthy	3.374354561	3	2

5. Classification

To validate our results, we have chosen the CNN (Convolutional Neural Network) to classify the diseases. CNN is a deep learning model that is often used in image classification. It comprises four layers: the Convolutional layer, the Pooling layer, the Activation function layer, and the Fully connected layer (Lu et al., 2021).

We have applied the CNN model twice to check whether our method has successfully segmented disease tissues in the leaf plant. First, the CNN was applied to the raw segmented version of the PlantVillage dataset without performing image segmentation on the disease leaf parts. Second, we have clustered the preprocessed dataset using the K-mean clustering algorithm with the optimal k value obtained by the Kneedle algorithm. Then we ran the CNN model on the processed dataset. To evaluate the performance of our model, the quantitative metric Accuracy (El-Hadj Imorou, 2020) was calculated using the following formulation:

$$\text{Accuracy (ACC)} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Where TP: true positive; FP: false positive, TN: true negative; and FN: false negative.

The train and test accuracy were recorded on both runs and are presented in Fig. 6. Our proposed model CNN result is apparent in Figure 6, left panel. Our model's CNN result achieved a training accuracy of 93.79% and a testing accuracy of 92.0% over 10 epochs. In contrast, the CNN result on the raw segmented dataset is presented in Figure 6, right panel. It achieved a training accuracy of 86.44 % and a testing accuracy of 80.46%.

Our proposed model showed a superior performance approach and proved that the proposed model is robust.

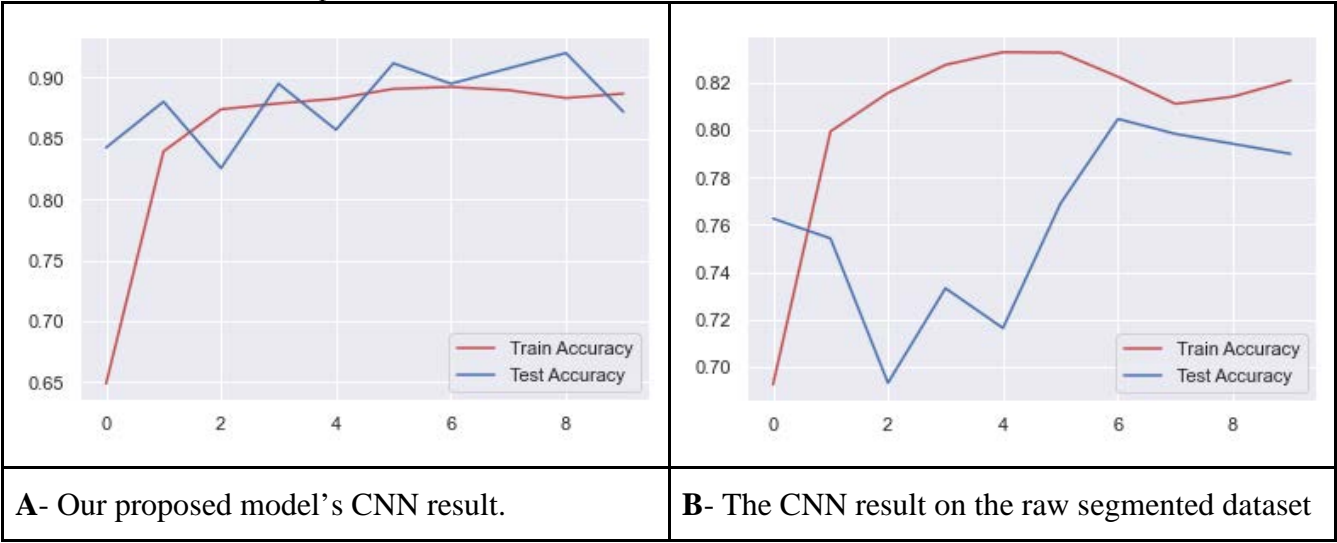


Fig. 6. The classification results of our proposed model are presented in the left panel, and the CNN results of the raw segmented dataset are presented in the right panel. The x-axis is the number of epochs, and the y-axis is the accuracy achieved.

6. Conclusion

This paper presents an artificial intelligence solution for detecting and classifying different plant leaf diseases, using the convolutional neural network for classification purposes. The segmentation of the plant leaf diseases is done using the k-means clustering algorithm. The optimal k value is obtained from the Kneedle algorithm, which showed superiority over the Elbow and the Silhouette method. Ambiguity is raised from the Elbow method while picking the optimal values of k. Besides, all the Silhouette plots with the three distance metrics displayed a zero measure, meaning that clusters are neutral; therefore, this method could not correctly distinguish clusters’.

Several disease plants are detected like black rot, rust, bacterial spot, early blight, late blight, leaf scorch, target spot, and mosaic viruses of different crops like apple, potato, tomato, grape, strawberry, and corn. The proposed hybrid model showed high accuracy in disease identification compared with the traditional k-means clustering algorithm.

Author contribution statements

All authors have contributed equally to this work. All authors discussed the results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Supplementary information

Accompanies this paper.

References:

- Anand, R., Veni, S., & Aravinth, J. (2016). An application of image processing techniques for detection of diseases on brinjal leaves using k-means clustering method. *2016 International Conference on Recent Trends in Information Technology (ICRTIT)*. <https://doi.org/10.1109/ICRTIT.2016.7569531>
- Arivazhagan, S., Shebiah, R. N., Ananthi, S., & Varthini, S. V. (2013). Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features. *Agricultural Engineering International: CIGR Journal*, 15(1), 211–217. <https://cigrjournal.org/index.php/Ejournal/article/view/2338>
- Bora, D. J., & Gupta, A. K. (2015). *A Novel Approach Towards Clustering Based Image Segmentation* (arXiv:1506.01710). arXiv. <https://doi.org/10.48550/arXiv.1506.01710>
- Cam, L. M. L., & Neyman, J. (1967). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification*. University of California Press.
- Craw, S. (2010). Manhattan Distance. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 639–639). Springer US. https://doi.org/10.1007/978-0-387-30164-8_506
- Cui, H., Wei, D., Ma, K., Gu, S., & Zheng, Y. (2021). A Unified Framework for Generalized Low-Shot Medical Image Segmentation With Scarce Data. *IEEE Transactions on Medical Imaging*, 40(10), 2656–2671. <https://doi.org/10.1109/TMI.2020.3045775>
- Dattorro, J. (2005). *Convex Optimization & Euclidean Distance Geometry*. Meboo Publishing USA.
- Devaraj, A., Rathan, K., Jaahnavi, S., & Indira, K. (2019). Identification of Plant Disease using Image Processing Technique. *2019 International Conference on Communication and Signal Processing (ICCSP)*, 0749–0753. <https://doi.org/10.1109/ICCSP.2019.8698056>
- El-Hadj Imorou, S. (2020). Socio-Economic and Health Determinants of Rural Households Consent to Prepay for Their Health Care in N'Dali (North of Benin). *Open Journal of Social Sciences*, 08(05), 348–360. <https://doi.org/10.4236/jss.2020.85024>
- Febrinanto, F., Dewi, C., & Triwiratno, A. (2019). The Implementation of K-Means Algorithm as Image Segmenting Method in Identifying the Citrus Leaves Disease. *IOP Conference Series: Earth and Environmental Science*. <https://doi.org/10.1088/1755-1315/243/1/012024>
- Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., Zhang, R., Hartmann, B. M., Zaslavsky, E., Sealfon, S. C., Chasman, D. I., FitzGerald, G. A., Dolinski, K., Grosser, T., & Troyanskaya, O. G. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 47(6), 569–576. <https://doi.org/10.1038/ng.3259>
- Gupta, M. K., & Chandra, P. (2021). Effects of similarity/distance metrics on k-means algorithm with respect to its applications in IoT and multimedia: A review. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-021-11255-7>
- Han, C.-Y. (2017). Improved SLIC image segmentation algorithm based on K-means. *Cluster Computing*, 20(2), 1017–1023. <https://doi.org/10.1007/s10586-017-0792-9>
- Hassan, S. M., Maji, A. K., Jasiński, M., Leonowicz, Z., & Jasińska, E. (2021). Identification of Plant-Leaf Diseases Using CNN and Transfer-Learning Approach. *Electronics*, 10(12), 1388. <https://doi.org/10.3390/electronics10121388>
- Hughes, D. P., & Salathe, M. (2016). *An open access repository of images on plant health to enable the development of mobile disease diagnostics* (arXiv:1511.08060). arXiv. <https://doi.org/10.48550/arXiv.1511.08060>
- Humaira, H., & Rasyidah, R. (2020). *Determining The Appropriate Cluster Number Using Elbow Method for K-Means Algorithm*. <https://doi.org/10.4108/eai.24-1-2018.2292388>
- Jain, A., Sarsaiya, S., Wu, Q., Lu, Y., & Shi, J. (2019). A review of plant leaf fungal diseases and its environment speciation. *Bioengineered*, 10(1), 409–424. <https://doi.org/10.1080/21655979.2019.1649520>
- Jain, S., Rejathalal, V., & Govindan, V. (2015). Image Segmentation using Sparse Subspace Clustering. *Undefined*. <https://www.semanticscholar.org/paper/Image-Segmentation-using-Sparse-Subspace-Clustering-Jain-Rejathalal/d657074d9b969f21024e46767162fca4b5b0bc34>
- Khalid, S., Ahmed, S., & Salman, M. (2020). A Novel Food Image Segmentation Based on Homogeneity Test of K-Means Clustering. *IOP Conference Series: Materials Science and Engineering*, 928, 032059. <https://doi.org/10.1088/1757-899X/928/3/032059>
- Li, X.-Y., Yu, L.-Y., Lei, H., & Tang, X.-F. (2017). The parallel implementation and application of an improved K-means algorithm. *Dianzi Keji Daxue Xuebao/Journal of the University of Electronic Science and Technology of China*, 46, 61–68. <https://doi.org/10.3969/j.issn.1001-0548.2017.01.010>
- Lu, J., Tan, L., & Jiang, H. (2021). Review on Convolutional Neural Network (CNN) Applied to Plant Leaf Disease Classification. *Agriculture*, 11(8), 707. <https://doi.org/10.3390/agriculture11080707>
- mariele. (2020, February 7). *How Aid Can Help Agriculture In Palestine*. Anera. <https://www.anera.org/blog/how-aid-can-help-agriculture-in-palestine/>
- Munisami, T., Ramsurn, M., Kishnah, S., & Pudaruth, S. (2015). Plant Leaf Recognition Using Shape Features and Colour Histogram with K-nearest Neighbour Classifiers. *Procedia Computer Science*, 58, 740–747. <https://doi.org/10.1016/j.procs.2015.08.095>
- Nanjundan, S., Sankaran, S., Arjun, C. R., & Anand, G. P. (n.d.). *Identifying the number of clusters for K-Means: A Hypersphere density based approach*. 5.
- Prakash, R., Saraswathy, G. P., Ramalakshmi, G., Mangaleswari, K. H., & Kaviya, T. (2017). Detection of leaf diseases and classification using digital image processing. *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. <https://doi.org/10.1109/ICIIECS.2017.8275915>

Ramesh, S., & Vydeki, D. (2020). Recognition and classification of paddy leaf diseases using Optimized Deep Neural network with Jaya algorithm. *Information Processing in Agriculture*, 7(2), 249–260. <https://doi.org/10.1016/j.inpa.2019.09.002>

Rangarajan, A. K., & Purushothaman, R. (2020). Disease Classification in Eggplant Using Pre-trained VGG16 and MSVM. *Scientific Reports*, 10. <https://doi.org/10.1038/s41598-020-59108-x>

Rousseeuw, P. (1987). Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Comput. Appl. Math.* 20, 53-65. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)

S, S., & Raghavendra, B. (2019). Diseases Detection of Various Plant Leaf Using Image Processing Techniques: A Review. *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*. <https://doi.org/10.1109/ICACCS.2019.8728325>

Satopaa, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a “Kneedle” in a Haystack: Detecting Knee Points in System Behavior. *2011 31st International Conference on Distributed Computing Systems Workshops*, 166–171. <https://doi.org/10.1109/ICDCSW.2011.20>

Sharma, P., Hans, P., & Gupta, S. C. (2020). Classification Of Plant Leaf Diseases Using Machine Learning And Image Preprocessing Techniques. *2020 10th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, 480–484. <https://doi.org/10.1109/Confluence47617.2020.9057889>

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, 2021(1), 31. <https://doi.org/10.1186/s13638-021-01910-w>

Shukla, S. (n.d.). *A Review ON K-means DATA Clustering APPROACH*. 14.

Sreedhar, C., Kasiviswanath, N., & Chenna Reddy, P. (2017). Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop. *Journal of Big Data*, 4(1), 27. <https://doi.org/10.1186/s40537-017-0087-2>

Starczewski, A., & Krzyżak, A. (2015). Performance Evaluation of the Silhouette Index. In L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh, & J. M. Zurada (Eds.), *Artificial Intelligence and Soft Computing* (pp. 49–58). Springer International Publishing. https://doi.org/10.1007/978-3-319-19369-4_5

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, 336, 012017. <https://doi.org/10.1088/1757-899X/336/1/012017>

Vadivel, A., Sural, S., & Majumdar, A. (2005). Human color perception in the HSV space and its application in histogram generation for image retrieval. *IS&T/SPIE Electronic Imaging*. <https://doi.org/10.1117/12.586823>

Yang, W., Cai, L., & Wu, F. (2020). Image segmentation based on gray level and local relative entropy two dimensional histogram. *PLOS ONE*, 15(3), e0229651. <https://doi.org/10.1371/journal.pone.0229651>

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, 2(2), 226–235. <https://doi.org/10.3390/j2020016>

Supplementary data

Table 1: The Silhouette method results for three distance metrics, Euclidean, Manhattan, and Cosine. The optimal k value and execution time are specified for each metric.

Folder Name	Attributes	Silhouette Method					
		Distance in Euclidean		Distance in Manhattan		Distance in Cosine	
		k value	Execution Time	k value	Execution Time	k value	Execution Time
Apple___Apple_scab	630	3	2min 53s	2	4min 38s	2	2min 27s
Apple___Black_rot	621	4	2min 49s	3	4min 36s	4	2min 22s
Apple___Cedar_apple_rust	275	5	1min 24s	4	1min 36s	4	1min 12s
Apple___healthy	1645	3	7min 38s	3	22min 25	3	6min 26s

Blueberry___healthy	1502	5	7min	4	19min 3s	3	5min 48s
Cherry_(including_sour)___healthy	854	3	3min 33s	3	7min 23s	3	3min 26s
Cherry_(including_sour)___Powdery_mildew	1052	3	4min 18s	3	10min 28s	3	3min 53s
Corn_(maize)___Cercospora_leaf_spot Gray_leaf_spot	513	2	2min 11s	2	3min 37s	2	1min 57s
Corn_(maize)___Common_rust	1192	2	5min 7s	2	13min 21s	2	4min 29s
Corn_(maize)___healthy	1162	2	4min 54s	2	13min 4s	2	4min 18s
Corn_(maize)___Northern_Leaf_Blight	985	3	4min 8s	3	9min 48s	3	3min 37s
Grape___Black_rot	1180	2	4min 46s	2	12min 58s	2	4min 31s
Grape___Esca_(Black_Measles)	1384	2	5min 44s	2	17min 1s	2	5min 25s
Grape___healthy	423	2	1min 37s	2	2min 31s	2	1min 35s
Grape___Leaf_blight_(Isariopsis_Leaf_Spot)	1076	2	4min 14s	2	11min 1s	2	4min 6s
Orange___Haunglongbing_(Citrus_greening)	5507	2	27min 26s	2	3h 27min 36s	2	27min 7s
Peach___Bacterial_spot	2297	2	8min 46s	2	40min 14s	3	8min 18s
Peach___healthy	360	4	1min 22s	5	2min 4s	3	1min 16s
Pepper,_bell___Bacterial_spot	997	3	3min 49s	3	9min 37s	3	3min 33s
Pepper,_bell___healthy	1478	3	5min 44s	3	18min 44s	3	5min 30s
Potato___Early_blight	1000	2	3min 42s	2	10min 27s	2	3min 28s
Potato___healthy	152	3	35.3 s	3	46.2 s	3	34.3 s
Potato___Late_blight	1000	2	3min 52s	2	10min 31s	2	3min 37s
Raspberry___healthy	371	3	1min 31s	3	2min 28s	4	1min 26s
Soybean___healthy	5090	3	24min 30s	3	3h 3s	3	22min 49s
Squash___Powdery_mildew	1835	2	7min 31s	2	27min 40s	2	7min 6s
Strawberry___healthy	456	2	1min 50s	2	3min 3s	2	1min 44s
Strawberry___Leaf_scorch	1109	2	4min 14s	2	11min 38s	2	4min 3s
Tomato___Bacterial_spot	2127	2	9min 8s	2	36min 23s	2	9min 1s
Tomato___Early_blight	1000	2	4min 10s	2	9min 52s	2	3min 47s
Tomato___healthy	1591	2	6min 2s	2	21min 19s	2	5min 44s
Tomato___Late_blight	1909	2	7min 44s	2	30min	2	7min 16s
Tomato___Leaf_Mold	952	2	3min 34s	2	9min 9s	2	3min 21s
Tomato___Septoria_leaf_spot	1771	2	6min 56s	2	25min 51s	2	7min 7s
Tomato___Spider_mites Two-spotted_spider_mite	1676	2	6min 58s	2	24min 2s	2	6min 34s
Tomato___Target_Spot	1404	2	5min 30s	2	16min 29s	2	5min 21s

Tomato___Tomato_mosaic_virus	373	4	1min 33s	5	2min 46s	4	1min 30s
Tomato___Tomato_Yellow_Leaf_Curl_Virus	5357	2	29min	2	3h 24min 47s	2	27min 46s

Table 2: The Kneedle Algorithm results

Folder	kneedle Algorithm	
	k value	mean k value
Apple___Apple_scab	3.487301587	3
Apple___Black_rot	3.325281804	3
Apple___Cedar_apple_rust	3.512727273	4
Apple___healthy	3.5556231	4
Blueberry___healthy	3.7310253	4
Cherry_(including_sour)___healthy	2.791569087	3
Cherry_(including_sour)___Powdery_mildew	3.975285171	4
Corn_(maize)___Cercospora_leaf_spot Gray_leaf_spot	4.040935673	4
Corn_(maize)___Common_rust_	4.651006711	5
Corn_(maize)___healthy	3.374354561	3
Corn_(maize)___Northern_Leaf_Blight	3.837563452	4
Grape___Black_rot	3.442372881	3
Grape___Esca_(Black_Measles)	3.562861272	4
Grape___healthy	3.122931442	3
Grape___Leaf_blight_(Isariopsis_Leaf_Spot)	3.894981413	4
Orange___Haunglongbing_(Citrus_greening)	3.386417287	3
Peach___Bacterial_spot	3.503265128	4
Peach___healthy	3.916666667	4
Pepper,_bell___Bacterial_spot	3.862587763	4
Pepper,_bell___healthy	3.48782138	3
Potato___Early_blight	4.007	4
Potato___healthy	3.401315789	3
Potato___Late_blight	4.121	4
Raspberry___healthy	3.078167116	3
Soybean___healthy	3.316222348	3
Squash___Powdery_mildew	4.043051771	4
Strawberry___healthy	3.728070175	4
Strawberry___Leaf_scorch	4.253381425	4
Tomato___Bacterial_spot	3.160789	3
Tomato___Early_blight	3.619	4
Tomato___healthy	4.083595223	4
Tomato___Late_blight	3.626506	4

Tomato__Leaf_Mold	3.719537815	4
Tomato__Septoria_leaf_spot	3.573687182	4
Tomato__Spider_mites Two-spotted_spider_mite	3.702863962	4
Tomato__Target_Spot	3.688746439	4
Tomato__Tomato_mosaic_virus	3.179624	3
Tomato__Tomato_Yellow_Leaf_Curl_Virus	3.62124323	4