SEMARAK ILMU
PUBLISHING
202103268166(003316878-P)

Semarak International Journal of Machine Learning

Journal homepage:
https://semarakilmu.my/index.php/sijml/index
ISSN: 3030-5241

Application of Different Distance Metrics on K-Means Clustering Algorithm for Retinal Vessel Images

Nor' Awatif Amri Muhammad Nazim[1], Normi Abdul Hadi[1], Mohd Rijal Ilias[1], Dian Kurniasari[3], Suhaila Abd Halim[1,2,*]

[1] School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA. 40450 Shah Alam, Selangor, Malaysia
[2] Smart Manufacturing Research Institute (SMRI), Universiti Teknologi MARA. 40450 Shah Alam. Selangor, Malaysia
[3] Department of Mathematics, Faculty of Science and Mathematics, Lampung University, Indonesia

ARTICLE INFO

Article history:
Received 5 November 2024
Received in revised form 30 November 2024
Accepted 20 December 2024
Available online 31 December 2024

ABSTRACT

Accurate segmentation of retinal blood vessels is important for the early detection and treatment of a variety of ocular disorders, including diabetic retinopathy and glaucoma. There are various methods used in image segmentation and one of them is K-means clustering. The problems of K-means clustering are its initial cluster centres, the spherical clusters' assumption, and the hard assignment of the pixels to the clusters, which has led to the improvements of the algorithm. These problems are closely related to the choice of distance metrics. In this study, the following objectives have been set that are to implement different distance metrics that are Euclidean, Manhattan, Chebychev and Mahalanobis distances in the K-means clustering algorithm to enhance retinal blood vessel segmentation and to measure the performance of the algorithms using accuracy, precision, recall, Dice Similarity coefficients (DSCs) and Jaccard similarity coefficients (JSCs). The retinal images are processed by choosing the green channel as it shows better visuals. Contrast Limited Adaptive Histogram Equalisation (CLAHE) is applied in the next step, followed by hole filling in order to improve the quality of the image. Next, we segment the blood vessels using K-means clustering. We apply each distance measurement separately to evaluate its impact on segmentation performance. We evaluate the segmentation algorithms' performance using ground truth and quantitative metrics. We implement the process using MATLAB. The results indicate that the choice of the distance metric significantly affects the segmentation accuracy. The Mahalanobis distance provides the best-balanced results between accuracy, precision, recall, Dice Similarity coefficient (DSC) and Jaccard similarity coefficient (JSC). Based on the findings, it is recommended to use Mahalanobis distance for optimal segmentation performance for retinal blood vessel images as it is suitable in identifying complex structures of vessels in the retinal.

Keywords:
K-Means clustering; Distance Metrics; MATLAB; Retinal Blood Vessels; image segmentation

* *Corresponding author.*
E-mail address: suhaila889@uitm.edu.my

https://doi.org/10.37934/sijml.4.1.1426a

1. Introduction

Clustering is one of the most popular techniques for image segmentation that divides the image pixels into different regions based on certain similarity criteria. It divides an image into several discrete regions to achieve high contrast and high similarity between them. It is an important tool in many fields, such as pattern identification, traffic picture processing, image processing, and health care. The methods exist for segmenting images, including cluster-based, neural network-based, threshold-based and edge-based approaches. The clustering method is one of the most effective approaches among the various techniques. Techniques such as subtractive clustering, fuzzy C-means clustering, mountain clustering and K-means clustering are widely used. Among these, K-means clustering is one of the most popular due to its simplicity and computational efficiency.

Compared to hierarchical clustering, K-means is not only simpler but also more computationally efficient, making it suitable for large datasets with many variables [1]. This efficiency is particularly beneficial when dealing with extensive image data, such as retinal blood vessel images. K-means clustering has a straightforward initial step. Start by figuring out how many clusters there are k and assuming that these clusters are centred. Any random objects can be chosen to be the initial centroids. K-means clustering is a widely used partitioning based clustering technique that has been successfully applied to a variety of real-world classification tasks in industry, business, and science. Its implementations aim to find a user-specified number of clusters that are represented by the centroids [2]. Many researchers manipulated the K-means algorithm to improve specific results. Zhao and Zhou [3] proposed new similarity calculations that are maximum and minimum distances known as the M-K-means algorithm and weighted Euclidean-distance known as W-K means algorithm that is applied to three sets of data.

Wiharto and Suryani [4] used K-means clustering in the analysis of retinal images to diagnose hypertensive retinopathy using computer aid. The listed algorithm was successful when conducting retinal blood vessel segmentation hence displaying differences between normal and hypertensive retinopathy conditions. The present study revealed that the selected number of clusters did affect fractal dimension values at the segmentation process.

Thomas and Kumar [5] explained a new approach to clustering of Magnetic Resonance Imaging of neurodisorder, which uses the Fuzzy C Means (FCM) algorithm with the Firefly Algorithm (FA). The work analyses the applicability of entropy measurements such as Renyi and Shannon entropy in the process of verifying a segmentation model that has been done using MATLAB 2020a. According to Shrivastava *et al.,* [6], K-means clustering algorithm is one of the main study methodologies in medical image segmentation by forming an important part in methodology. In addition, the K-means algorithm is applied to grayscale images which provide a good quality of output images [7].

Simhachalam and Ganesan [8] explained that K-means and Fuzzy c-mean (FCM) used the normalized Euclidean distance measure on the liver data set and wine data set. The K-means is established to be more efficient than FCM which had the best accuracy.

According to Miao *et al*., [9], one single plant segmentation via the method based on Euclidean clustering can efficiently separate maize plants with different planting densities at the V5 and V6 stages from another by utilizing the K-means clustering method. The performance of clustering resembles the Euclidean clustering which is suitable for plants with leaves arranged in a triangular manner that do not cross one another. While K-means clustering is suitable for plants with leaves that cross one another. In Tamilselvi *et al.,* [10], distance measures of the squared Euclidean and city-Block were used. The study presented a new colour-based K-means clustering algorithm for image segmentation with the purpose of detecting infected leaves within agricultural fields.

In clustering algorithms, the selection of a distance metric is crucial as it affects how data points are grouped into discrete clusters that show similarity between the data points, where it influences the clustering results. In deciding which distance metrics were most appropriate for retinal blood vessel segmentation, several factors were taken into consideration due to the morphological complexity of vascular structures in retinal images. Euclidean distance, being one of the most commonly used metrics, captures geometric relationships between points and is effective in grouping pixels based on their spatial proximity. This is crucial for maintaining the continuity of vessel structures. Manhattan distance measures the sum of absolute differences between points and is particularly useful for capturing the vertical and horizontal alignment of vessel structures, which are often crucial in medical images. The Chebychev distance measures the maximum difference along any coordinate dimension and is effective in identifying the extremities of the vessel structures, ensuring that the segmentation process captures the full extent of the vessels. The Mahalanobis distance accounts for correlations between variables, making it well-suited for complex datasets where the features are not independent, such as the intensity values of pixels in retinal images. By understanding the implications of using these metrics, the study aims to customize segmentation strategies to the complex patterns found in retinal images. Each distance metric offers unique advantages that can enhance the accuracy and robustness of the segmentation process. By examining the performance of K-means clustering with these different metrics, the study seeks to identify the most effective approach for retinal blood vessel segmentation. The tailored approach ensures that the specific characteristics of retinal images are addressed, leading to improved segmentation outcomes. Consequently, the selected distance metrics play a critical role in achieving accurate and reliable segmentation of retinal blood vessels, ultimately contributing to better diagnostic and treatment processes in medical imaging.

The retinal images often contain noise artifacts and varying contrast that significantly affect the results of clustering that leads to inaccurate evaluations. In addition to that, most studies focused on a single metric without exploring different metrics applied to K-means clustering [4,11-12]. Dash, Parida and Bhoi [11] applied Euclidean distance in K-mean clustering to extract blood vessel from Fundus images. Wiharto and Suryani [4] compared different clustering algorithms of K-means and Fuzzy c-means for segmentating retinal blood vessels. Both algorithms manipulated Euclidean distance to determine the membership of a cluster. Geetharamani and Balasubramanian [12] also applied Euclidean distance on K-mean clustering to segment blood vessel on STARE database and DRIVE database. The correct choice of distance metrics is required to ensure that the clustering of retinal blood vessels is accurately extracted for further processed.

The paper highlights the use of different distance metrics in k-means clustering and the selection of images that give the most efficient results in different image clusters. Hence the objectives of the paper are to implement four different metrics that are Euclidean distance, Manhattan distance, Chebychev distance and Mahalanobis distance on K-means clustering for the segmentation process. Then, the performance of the study was evaluated using accuracy, precision, recall, Dice Similarity coefficient (DSC) and Jaccard similarity coefficient (JSC).

2. Methodology

Figure 1 shows the general flowchart of the proposed algorithm of using K-means clustering on retinal blood vessel images. The flowchart begins with data collection of the images and ends with performance evaluation using accuracy, precision, recall, DSC and JSC. All the processes were implemented using MATLAB.

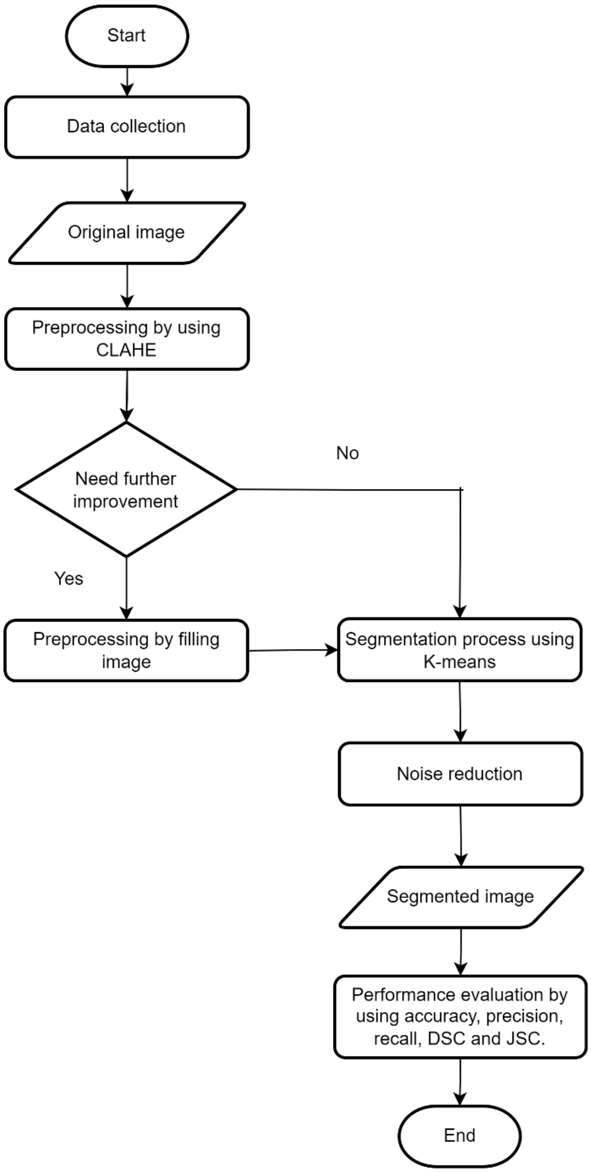


Fig. 1. General flowchart of proposed algorithm

2.1 Data Collection

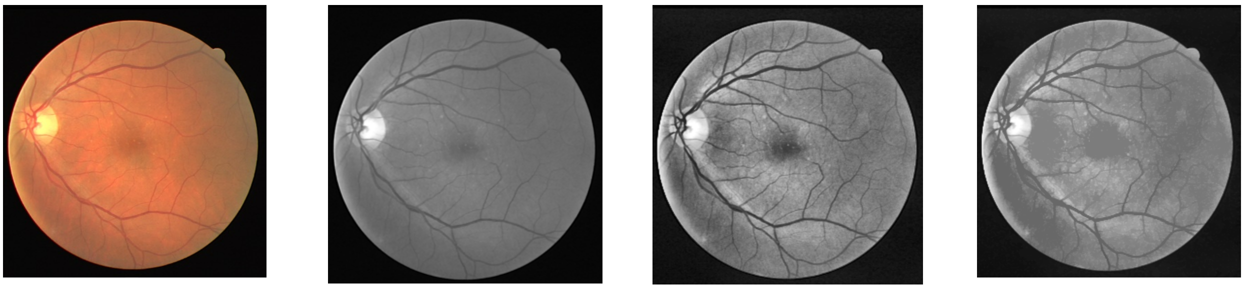
The data collection is focused on the use of the DRIVE (Digital Retinal Images for Vessel Extraction) database. Introduced by Staal *et al.,* [13], the DRIVE database is a widely recognized resource for researchers in the field of retinal image analysis, particularly vessel segmentation. The database consists of 40 colour JPEG fundus photos in which 7 of them had abnormal pathology cases. Twenty photos were allocated to the training set and twenty to the testing set. Due to time constraints, only 20 samples of retinal blood vessel images in JPEG were used in this study. Figure 2 shows the three samples of the retinal blood vessel images.

(a) Image 1 (b) Image 2 (c) Image 3

Fig. 2. Three samples of blood vessel images used in the study

2.1 Preprocessing

In the preprocessing stage, the image quality is improved for further processing. The color or RGB image consists of red, green and blue channels. In this study, the green channel of the RGB image is selected for further processing. The green channel is used because the channel gives higher sensitivity to the vessel structures. Figure 3 shows the contrast of the extracted green channel image that is enhanced using contrast limited adaptive histogram equalization (CLAHE). The green channel is used because microaneurysms are in high contrast and clearly visible in green channel [14]. CLAHE is a method that adjusts the contrast of the image based on the intensity distribution in localized regions, which improves the visibility of blood vessel features. The localised approach improves the contrast more efficiently in regions with diverse intensity distributions compared to conventional histogram equalisation techniques, which use a universal transformation [15]. The CLAHE process begins with determining the region size and clip limit, followed by shaping the histogram for each region, distributing excess clip limit values, mapping the new histogram to the image, and interpolating pixels in neighbouring regions to generate the final CLAHE image [16]. The default clip limit is used to provide a good balance between contrast enhancement and noise suppression for general use cases. The result of each enhancement process is shown in Figure 3.



(a) Original Image 1 (b) Converted Image 1 (c) Enhanced Image 1 (d) Filled Image 1

Fig. 3. Sample results of Image 1

Morphological operations such as hole filling were applied to remove any small light spots within dark regions, which are not part of the retinal vessels. We filled them with the 'holes' argument. This process is effective for removing small light regions from larger dark regions, which can aid in succeeding the segmentation step.

2.2 Segmentation

In this study, the K-means algorithm was used to find out the groups/clusters in the data where the value k is the desired number of groups. The algorithm iteratively assigns each data points to one of k groups, considering the provided features [17].

It also takes input parameter k (number of clusters) and a set of n objects that need to be partitioned into some number k of clusters to increase the intra-cluster similarity and minimize inter-cluster similarity. The main idea is to divide the into k number of centroids, k denotes a cluster. These centroids are evenly distasteful as their place performs a vital influence on the outputs. The best location would be where the centroids are farthest apart from each other. Afterward, every data point in the data set is connected to its closest centroid. After doing this one time, k new centroids were re-calculated. This process was repeated all over again along with a new assignment of data points to the nearest centroid. This loop will continue until the position of the centroids no more change, and this indicates that the algorithm has converged.

The K-means Algorithm [18]:

i. Choose k random points s cluster centers.
ii. Assign each observation to the nearest cluster center.
iii. Calculate new clusters.
iv. If the clusters are changes, go to step 2, Otherwise, stop.

Four different distance metrics were used in this study that are Euclidean distance, Manhattan distance, Chebychev distance and Mahalanobis distance. The properties of each metric are different and decide how the pixels' similarity is determined, thus affecting the clustering result. Most machine learning algorithms, including K-means used the distance metric to measure the similarity between observations.

2.2.1 Distance Metrics

In identifying the distance metrics to be used in K-means clustering, appropriate metrics are required because they affect the convergence and identification of cluster centroids. Each metric highlights different aspects of data, perhaps resulting in various cluster shapes and sizes.

Euclidean distance is the most commonly used metric for a k-dimensional space. It calculates the distance between two points in a straight line. The formula for the Euclidean distance is given as:

$$d_e = \sqrt{\sum_{i=1}^{k}\left(x_i - y_i\right)^2} \tag{1}$$

where, k is the number of dimensions and x_i and y_i are the coordinates or feature values of the two points being compared in the i-th dimension.

Manhattan distance is the sum of absolute differences between points across all the dimensions. It is also known as city block distance. In order to calculate Manhattan distance, the sum of absolute distances in both the x and y directions were set. The generalized formula of Manhattan distance for a k-dimensional space is given as:

$$d_{man} = \sum_{i=1}^{k} | x_i - y_i | \tag{2}$$

Chebychev distance, which is often referred to as maximum value distance, is calculated as the absolute magnitude of the variations in the coordinates of two objects [19].

$$d_b = \max_i \{|x_i - y_i|\} \tag{3}$$

K-means can be used to identify peaks in a Gaussian distribution but the shape of the area around that peak or cluster that gets detected depends on the distance metrics used. Like explained earlier, Euclidean distance is used for the formation of spherical or circular clusters, Manhattan for rhombus shaped clusters, Chebychev for square clusters. Mahalanobis distance metric similarly forms elliptical clusters and is defined as the statistical distance between two points in the p-dimensional space calculated by the formula [20].

$$d_{mah} = \sqrt{(x_i - y_i) C^{-1} (x_i - y_i)'} \tag{4}$$

where C is the covariance matrix.

2.3 Noise Reduction

After the initial segmentation using K-means clustering, post-processing techniques are applied to refine the segmented blood vessels and reduce any remaining noise. Morphological operations such as erosion and dilation are used to remove small, isolated noise pixels and to fill gaps in the segmented blood vessels. This step ensures that the segmented blood vessels are clean and accurately represent the actual vascular structures in the retina [21].

2.4 Performance Evaluation

Comparing with the ground truth of the images that illustrating the actual position of blood vessels is critical importance as it determines the degree of accuracy in clustering results. The segmented results based on proposed algorithm are measured using accuracy, precision, recall, DSC and JSC that present clear quantitative results of the clustering algorithm. These measurements are used because the metrics provide robust evaluation. High score indicates that the clustering method is accurate in their detection.

The performance evaluations provide the information regarding the quality of the algorithm in terms of the identification of the vessels and coverage of all vessel structures. The notation of the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) represents the outcomes of the clustering that can be analysed in detail. The accuracy represents the percentage of correctly classified among all the object in dataset [22]. The DSC, also known as the F1 Score in some contexts. It is a measure of similarity between the predicted and ground truth segments. The JSC, also known as Intersection over Union (IoU), measures the ratio of the intersection of the predicted and actual sets to their union [23].

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{5}$$

TP: The total number of identified vessel pixels correctly classified.
TN: The total number of non-vessel pixels that are correctly classified.
FP: The total number of non-vessel pixels incorrectly identified as vessel pixels.
FN: The total number of vessel pixels that were falsely classified as non-vessel pixels.

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{6}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{7}$$

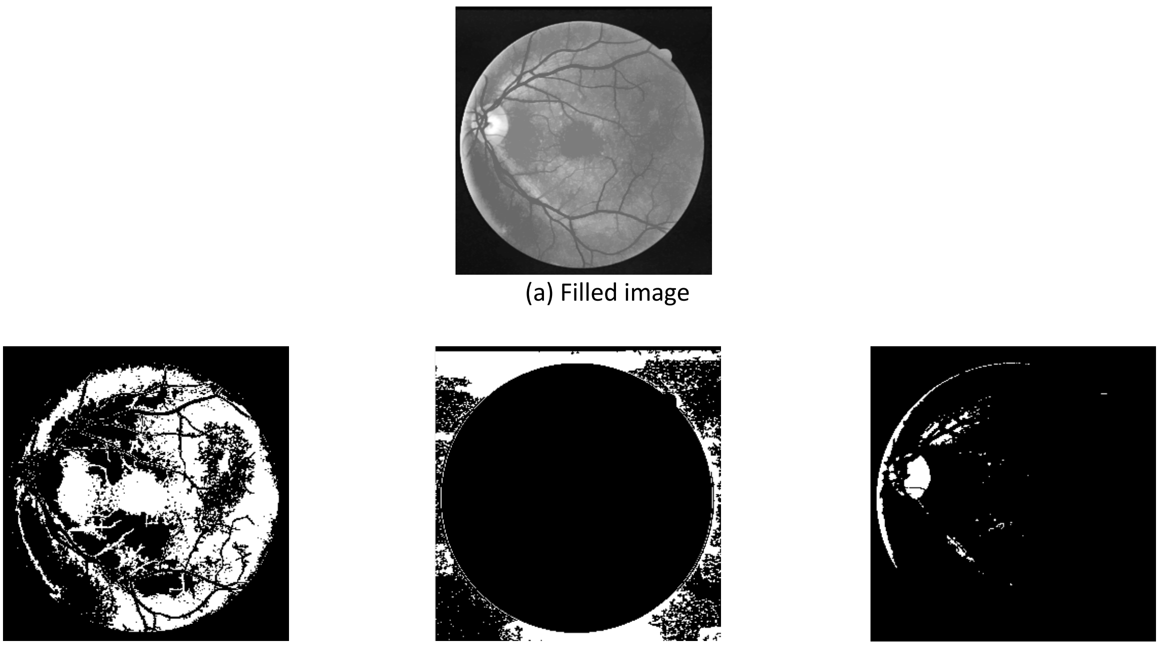
$$\text{DSC}=\frac{\text{2TP}}{\text{2TP+FP+FN}} \tag{8}$$

$$\text{JSC}=\frac{\text{DSC}}{\text{2-DSC}} \tag{9}$$

Analysing the study by Muller *et al.,* [24], certain points are important in performance evaluation measures in the case of medical picture segmentation which includes the DSC and JSC.

3. Results

Generally, in this study the images were clustered into 6 clusters. The results of the clusters in K-means clustering using Euclidean as the segmentation process that was applied to Image 1 in this study are shown in Figure 4.



(a) Filled image

(b) Cluster 1 (c) Cluster 2 (d) Cluster 3

(e) Cluster 4 (f) Cluster 5 (g) Cluster 6

Fig. 4. Six clusters of Image 1

Each cluster gives different segmentation results. The selection of cluster is important to ensure the selected area was segmented accurately. It can be observed that for each cluster, cluster 6 (g) shows the most features of the segmented retinal vessels. Hence, Image 1 with cluster 6 is chosen for further analysis.

Figure 5 shows the results of the selected cluster on distance metrics on K-means clustering using that was applied on Image 1 in this study.

(a) Euclidean Distance (Cluster 6) (b) Manhattan (Cluster 2) (c) Chebychev (Cluster 6) (d) Mahanalobis (Cluster 6)

Fig. 5. Different Results of Different Distance Metric for Image 1

For all the 20 images, the segmented images for each cluster were manually observed and collected. In addition to that, the quantitative measurements for the results are measured and compared.

The performance of each distance metric on the quality of the resulting segments was evaluated using accuracy, precision, recall and the DSC and JSC. Based on the results obtained, it is possible to state that the range of distance metrics influences the general characteristics of segmentation and clustering.

Table 1 presents a detailed performance evaluation with the best distance metric (red color) for 20 blood vessel images that focuses specifically on accuracy, precision, recall, DSC and JSC. These evaluations help in identifying the most effective distance metric for accurate and precise segmentation of retinal blood vessel images.

The Manhattan distance consistently demonstrated superior performance in most cases. For example, in the segmentation in Image 1, the Manhattan distance achieved the highest accuracy (0.910**)** and recall (0.633) compared to Euclidean, Chebyshev and Mahalanobis distances. The Manhattan distance excels in capturing the fine details and variations in retinal blood vessels, making it a robust choice for this application.

Table 1
Performance of each distance metric for all the 20 sample of Retinal blood vessel images

	Accuracy				Precision				Recall				DSC				JSC			
Image	d_e	d_{man}	d_c	d_{mah}	d_e	d_{man}	d_c	d_{mah}	d_e	d_{man}	d_c	d_{mah}	d_e	d_{man}	d_c	d_{mah}	d_e	d_{man}	d_c	d_{mah}
1	0.605	0.910	0.906	0.906	0.000	0.432	0.055	0.055	0.001	0.633	0.016	0.016	0.000	0.514	0.025	0.025	0.000	0.346	0.013	0.013
2	0.752	0.781	0.732	0.895	0.072	0.045	0.083	0.428	0.148	0.071	0.196	0.481	0.097	0.055	0.117	0.453	0.051	0.029	0.062	0.293
3	0.618	0.819	0.883	0.915	0.005	0.021	0.027	0.376	0.022	0.039	0.022	0.435	0.007	0.028	0.024	0.403	0.004	0.014	0.012	0.253
4	0.748	0.565	0.848	0.896	0.040	0.001	0.041	0.585	0.052	0.002	0.014	0.339	0.045	0.001	0.021	0.429	0.023	0.001	0.011	0.273
5	0.896	0.734	0.892	0.905	0.113	0.074	0.434	0.510	0.012	0.154	0.430	0.244	0.022	0.100	0.432	0.330	0.011	0.053	0.276	0.198
6	0.817	0.814	0.862	0.753	0.271	0.268	0.038	0.060	0.701	0.706	0.027	0.134	0.391	0.389	0.032	0.083	0.243	0.241	0.016	0.043
7	0.863	0.912	0.900	0.925	0.342	0.500	0.111	0.649	0.603	0.093	0.019	0.319	0.436	0.157	0.033	0.428	0.279	0.085	0.017	0.272
8	0.878	0.772	0.865	0.919	0.059	0.067	0.371	0.650	0.016	0.103	0.550	0.367	0.026	0.081	0.443	0.469	0.013	0.042	0.284	0.306
9	0.753	0.595	0.879	0.902	0.069	0.001	0.348	0.086	0.155	0.005	0.507	0.017	0.095	0.002	0.412	0.028	0.050	0.001	0.260	0.014
10	0.731	0.810	0.793	0.605	0.000	0.230	0.037	0.008	0.001	0.604	0.065	0.032	0.000	0.333	0.047	0.013	0.000	0.200	0.024	0.006
11	0.622	0.798	0.915	0.933	0.005	0.153	0.039	0.440	0.028	0.518	0.017	0.437	0.009	0.236	0.024	0.438	0.004	0.134	0.012	0.281
12	0.916	0.783	0.876	0.748	0.433	0.097	0.343	0.000	0.088	0.201	0.561	0.000	0.147	0.131	0.426	0.000	0.079	0.070	0.270	0.000
13	0.846	0.906	0.600	0.869	0.028	0.425	0.001	0.334	0.027	0.471	0.003	0.626	0.027	0.447	0.001	0.435	0.014	0.288	0.001	0.278
14	0.824	0.819	0.824	0.782	0.242	0.237	0.172	0.105	0.375	0.382	0.208	0.162	0.294	0.293	0.188	0.127	0.172	0.171	0.104	0.068
15	0.892	0.876	0.883	0.871	0.408	0.365	0.045	0.041	0.543	0.574	0.018	0.022	0.466	0.446	0.025	0.029	0.304	0.287	0.013	0.015
16	0.905	0.846	0.590	0.799	0.576	0.064	0.002	0.061	0.466	0.030	0.006	0.059	0.515	0.041	0.003	0.060	0.347	0.021	0.002	0.031
17	0.592	0.818	0.830	0.898	0.001	0.265	0.279	0.401	0.002	0.611	0.600	0.335	0.001	0.369	0.381	0.365	0.000	0.227	0.236	0.223
18	0.593	0.882	0.757	0.744	0.000	0.385	0.032	0.090	0.001	0.624	0.062	0.215	0.000	0.476	0.042	0.127	0.000	0.312	0.022	0.068
19	0.710	0.776	0.919	0.933	0.067	0.028	0.534	0.693	0.183	0.047	0.492	0.403	0.098	0.035	0.512	0.509	0.052	0.018	0.344	0.342
20	0.743	0.619	0.695	0.911	0.025	0.002	0.074	0.431	0.064	0.009	0.262	0.523	0.036	0.004	0.115	0.472	0.018	0.002	0.061	0.309

The Mahalanobis distance also showed promising results, especially in images where complex vessel structures are present. For instance, in Image 2 and Image 4, Mahalanobis distance provided higher precision of 0.428 and 0.585 respectively and the DSC of 0.453 and 0.429 respectively. Suggesting its capability in handling variations in vessel thickness and orientation. This makes Mahalanobis distance a viable option for applications requiring high precision and detail. The performance demonstrates the Mahalanobis distance's ability to account for correlations in the data, which is critical for accurate segmentation. Unlike other distance metrics, which may struggle with data with variable degrees of correlation, the Mahalanobis distance considers feature covariance, allowing it to more reliably discern between blood vessel and non-vessel pixels. The results are more robust segmentation performance, as indicated by balanced metrics across several parameters.

The highest value of accuracy (0.933) for Image 19 gives the impact that the Mahalanobis distance is able to cluster the pixels to their correct groups. For the precision, and recall values shows not only that it correctly identifies blood vessels but also that it can consistently do so. This makes Mahalanobis distance versatile and since feature vessels are crucial in segmentation, precision and recall is achieved. Therefore, for such tasks where structure of data and their correlation are complex, Mahalanobis distance is effective measure of which leads to improvement of appearance of the overall performance of the K-means clustering algorithm in the segmentation of retinal blood vessel.

On the other hand, the Chebyshev distance exhibited mixed results. While it performed well in certain images, such as Image 5, where it achieved high accuracy (0.891) and precision (0.434), its overall performance was less consistent compared to Manhattan and Mahalanobis distances. This suggests that Chebyshev distance more suitable for specific types of images or conditions rather than as a general-purpose distance metric for retinal vessel segmentation. Its performance variability indicates that it is sensitive to certain image characteristics, such as noise or variations in vessel intensity.

Finally, the Euclidean distance, despite being the most used metric, did not consistently outperform the other distance metrics. Its performance varied significantly across different images, suggesting that while it is simple to implement and computationally efficient, it may not always be the best choice for complex segmentation tasks like retinal blood vessels. The variability in performance indicates that Euclidean distance might not be able to capture the intricate details and variations in retinal vessel structures as effectively as Manhattan or Mahalanobis distances. The analysis highlights the importance of selecting an appropriate distance metric based on the specific characteristics of the image data and the segmentation requirements.

The noise and artifacts in the images also affect the poor segmentation of blood vessels that are usually found in the images of the retinal. If the preprocessing steps are not suitable for removing unwanted noise or enhancing the salient features properly, proper clustering cannot be applied and lead to formation of irrelevant clusters.

4. Conclusions

In conclusion, the study aims to enhance the quality of the images of the digital retinal blood vessel images using CLAHE, building the K-means clustering method using the various distances and to evaluate the performance of the proposed algorithm in terms of accuracy, precision, recall, DSCs and JSCs. The objectives of the study were achieved to a satisfactory level due to the methodical approach using image preparation and segmentation. Applying CLAHE meant an increase in the difference and perceptibility of the retinal blood vessels. Therefore, subsequent more accurate segmentation. This improvement turned out to be very helpful especially in differentiating the blood vessels from the background especially when taking pictures with different lighting conditions.

Second, the use of the K-means clustering case detecting several measuring distances that are Euclidean, Manhattan, Chebychev and Mahalanobis distances that discover the chosen measure influences the segmentation result greatly. It was outlined that the selected distance measures had higher sensitivity and accuracy in terms of identifying certain structures of blood vessels, thus prove the effectiveness of the chosen methodology.

The evaluation of various distance metrics for retinal blood vessel segmentation using the K-means clustering algorithm reveals significant insights into their performance. The Mahalanobis distance emerged as the most effective metric, demonstrating its superior ability to handle data variability and distribution. This distance metric consistently achieved high accuracy, precision, recall, DSC and JSC values, indicating its robustness in accurately identifying and segmenting retinal blood vessels. Its ability to account for correlations in the data makes it particularly suitable for complex tasks where the structure and relationship between features are critical. The study underscores the importance of considering data characteristics and segmentation requirements when selecting distance metrics to ensure accurate and reliable segmentation outcomes.

For future investigations and practical utilization. future research should make efforts to include other methods of clustering and different models of segmentation to improve the effectiveness of the methodology. Also, there is a possibility to hybridize the work of deep learning methods and some classic clustering techniques can bring new results and enhance the analysis of the retinal images. The researchers can also extend the set of acquired retinal images and investigate the algorithm performance when tested with different conditions in order to improve the results applicability.

Acknowledgement
This research was not funded by any grant.

References

[1] Fahim, A. M., A. M. Salem, F. Af Torkey, and MA1101 Ramadan. "An efficient enhanced k-means clustering algorithm." *Journal of Zhejiang University-Science A* 7 (2006): 1626-1633. https://doi.org/10.1631/jzus.2006.A1626

[2] Kapil, Shruti, and Meenu Chawla. "Performance evaluation of K-means clustering algorithm with various distance metrics." In *2016 IEEE 1st international conference on power electronics, intelligent control and energy systems (ICPEICES)*, pp. 1-4. IEEE, 2016. https://doi.org/10.1109/ICPEICES.2016.7853264

[3] Zhao, YanPing, and XiaoLai Zhou. "K-means clustering algorithm and its improvement research." In *Journal of Physics: Conference Series*, vol. 1873, no. 1, p. 012074. IOP Publishing, 2021. https://doi.org/10.1088/1742-6596/1873/1/012074

[4] Wiharto, Wiharto, and Esti Suryani. "The comparison of clustering algorithms K-means and fuzzy C-means for segmentation retinal blood vessels." *Acta Informatica Medica* 28, no. 1 (2020): 42. https://doi.org/10.5455/aim.2020.28.42-47

[5] Thomas, Elisabeth, and S. N. Kumar. "Fuzzy C Means Clustering Coupled with Firefly Optimization Algorithm for the Segmentation of Neurodisorder Magnetic Resonance Images." *Procedia Computer Science* 235 (2024): 1577-1589. https://doi.org/10.1016/j.procs.2024.04.149

[6] Shrivastava, Kalpana, Neelesh Gupta, and Neetu Sharma. "Medical image segmentation using modified k means clustering." *International Journal of Computer Applications* 103, no. 16 (2014). https://doi.org/10.5120/18157-9341

[7] Deeparani, K., and P. Sudhakar. "Efficient image segmentation and implementation of K-means clustering." *Materials Today: Proceedings* 45 (2021): 8076-8079. https://doi.org/10.1016/j.matpr.2021.01.154

[8] Simhachalam, B., and G. Ganesan. "Performance comparison of fuzzy and non-fuzzy classification methods." Egyptian informatics journal 17, no. 2 (2016): 183-188. https://doi.org/10.1016/j.eij.2015.10.004.

[9] Miao, Yanlong, Shuai Li, Liuyang Wang, Han Li, Ruicheng Qiu, and Man Zhang. "A single plant segmentation method of maize point cloud based on Euclidean clustering and K-means clustering." *Computers and Electronics in Agriculture* 210 (2023): 107951. https://doi.org/10.1016/j.compag.2023.107951

[10] Tamilselvi, P., P. Nagasankar, and S. Geetha. "Color based K-Means Clustering For Image Segmentation to Identify the Infected Leaves." In *2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, vol. 1, pp. 41-48. IEEE, 2019. https://doi.org/10.1109/ICONSTEM.2019.8918737
[11] Dash, Jyotiprava, Priyadarsan Parida, and Nilamani Bhoi. "Retinal blood vessel extraction from fundus images using enhancement filtering and clustering." *ELCVIA: electronic letters on computer vision and image analysis* 19, no. 1 (2020): 0038-52. https://doi.org/10.5565/rev/elcvia.1239
[12] Geetharamani, R., and Lakshmi Balasubramanian. "Automatic segmentation of blood vessels from retinal fundus images through image processing and data mining techniques." *Sadhana* 40 (2015): 1715-1736. https://doi.org/10.1007/s12046-015-0411-5
[13] Niemeijer, M., J. J. Staal, Bv Ginneken, M. Loog, and M. D. Abramoff. "DRIVE: digital retinal images for vessel extraction." Methods for evaluating segmentation and indexing techniques dedicated to retinal ophthalmology (2004).
[14] Indumathi, G., and V. Sathananthavathi. "Microaneurysms detection for early diagnosis of diabetic retinopathy using shape and steerable Gaussian features." In *Telemedicine Technologies*, pp. 57-69. Academic Press, 2019. https://doi.org/10.1016/B978-0-12-816948-3.00005-2
[15] Venkatesh, S., C. John De Britto, P. Subhashini, and K. Somasundaram. "Image Enhancement and Implementation of CLAHE Algorithm and Bilinear Interpolation." *Cybern. Syst.* (2022): 1-13. https://doi.org/10.1080/01969722.2022.2147128
[16] Hana, F. M., and I. D. Maulida. "Analysis of contrast limited adaptive histogram equalization (CLAHE) parameters on finger knuckle print identification." In *Journal of Physics: Conference Series*, vol. 1764, no. 1, p. 012049. IOP Publishing, 2021. https://doi.org/10.1088/1742-6596/1764/1/012049
[17] Dhanachandra, Nameirakpam, Khumanthem Manglem, and Yambem Jina Chanu. "Image segmentation using K-means clustering algorithm and subtractive clustering algorithm." *Procedia Computer Science* 54 (2015): 764-771. https://doi.org/10.1016/j.procs.2015.06.090
[18] Hull, John C. "Machine learning in business: An introduction to the world of data science." *(No Title)* (2021).
[19] Singh, Archana, Avantika Yadav, and Ajay Rana. "K-means with Three different Distance Metrics." *International Journal of Computer Applications* 67, no. 10 (2013). https://doi.org/10.5120/11430-6785
[20] Chokniwal, Ankita, and Manoj Singh. "Faster Mahalanobis K-means clustering for Gaussian distributions." In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 947-952. IEEE, 2016. https://doi.org/10.1109/ICACCI.2016.7732167
[21] Raudonis, Vidas, Agne Paulauskaite-Taraseviciene, Kristina Sutiene, and Domas Jonaitis. "Towards the automation of early-stage human embryo development detection." *Biomedical engineering online* 18 (2019): 1-20. https://doi.org/10.1186/s12938-019-0738-y
[22] Mohamad, Muhammad Arif, and Muhammad Aliif Ahmad. "Handwritten Character Recognition using Enhanced Artificial Neural Network." *Journal of Advanced Research in Computing and Applications* 36, no. 1 (2024): 1-9. https://doi.org/10.37934/arca.36.1.19
[23] Ghani, Noor Ain Syazwani Mohd, Abdul Kadir Jumaat, Rozi Mahmud, Mohd Azdi Maasar, Farizuwana Akma Zulkifle, and Aisyah Mat Jasin. "Breast abnormality boundary extraction in mammography image using variational level set and self-organizing map (SOM)." *Mathematics* 11, no. 4 (2023): 976. https://doi.org/10.3390/math11040976
[24] M Müller, Dominik, Iñaki Soto-Rey, and Frank Kramer. "Towards a guideline for evaluation metrics in medical image segmentation." *BMC Research Notes* 15, no. 1 (2022): 210. https://doi.org/10.1186/s13104-022-06096-y