J. Electrical Systems 20-11s (2024): 499-510

Pradeep Kumar KG[1*]

Dr. Karunakara K[2]

Dr. Thyagaraju G.[3]

Detection of Diabetic Retinopathy Using Improvised Fuzzy Contextual Data Clustering Method

Journal of Electrical Systems

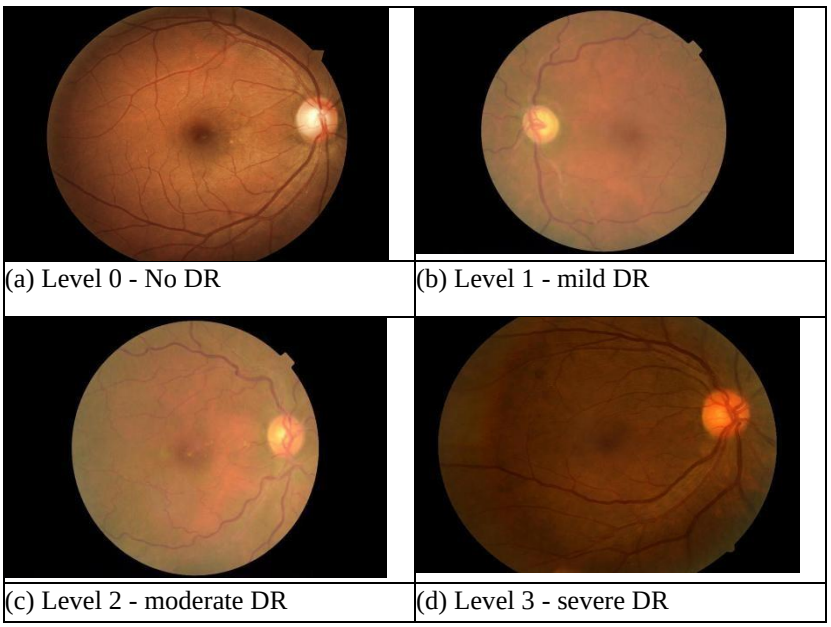
Abstract: The study introduces the vigorous fuzzy contextual data K-means clustering technique, an advanced version of the traditional k-means clustering method that incorporates localized information parameters customized for each cluster. A comparative analysis is performed between the robust fuzzy local information k-means clustering and the modified fuzzy C means clustering, which enhances Fuzzy C Means with a median adjustment parameter for diabetic retinopathy detection. The three datasets: IDRiD, DIATREB1, and DIATREB2 fundus images are used in this research. The proposed algorithm achieves a 94.4% accuracy rate. It is designed to efficiently categorize a large volume of retinal images, thereby improving performance and addressing the critical need for prompt and accurate diagnoses in diabetic retinopathy care.

Keywords - Clustering, Fuzzy technique, Deep neural networks, Fundus images, Machine learning

1. INTRODUCTION

Diabetic retinopathy or DR, the leading eye disease of diabetes, affects just about one-third of all individuals with diabetes around the globe and almost all of those living in areas of high prevalence of diabetes. In India where 60 to 65 million or 8 per cent of its population suffer from diabetes, DR affects between 13 and 20 per cent of its population. India and many such countries with comparable public health systems-challenges are in desperate need of more eye care facilities for the large number of citizens hurting from diabetes who will go on to develop eye diseases[1]. For the affected, there are significant obstacles to getting screened and treated for DR, especially issues of patient awareness and accessibility to screening. As well as the shortage of trained and qualified ophthalmologists and clinical teams. DR can eventually lead to significant loss of sight and is a primary cause of sightlessness among adults in many high-income countries. Therefore, routine eye examinations are essential for DR and the ADA(American Diabetes Association) recommends per-year dilated eye exams for those with diabetes. Research continues to highlight the growing need to better understand DR and the development of better treatments and eye screening for DR.

We identify DR by the presence of several types of lesions – microaneurysms (MA), hard exudates (EX) and haemorrhages (HM) – that manifest as red (for MA and HM) and bright (for soft and hard) lesions, or a mixture of both. DR is described in five stages supported on the existence of the lesions mentioned above: proliferative DR, severe DR, moderate DR, mild DR and no DR. Deep learning has achieved state-of-the-art results in numerous fields, but often in supervised settings where labelled data is required. However, labelled data is costly to acquire and label, which represents a large research opportunity to alleviate the burden of this task for real-time applications[2][3].



(a) Level 0 - No DR | (b) Level 1 - mild DR

(c) Level 2 - moderate DR | (d) Level 3 - severe DR

[1*]Vivekananda College of Engineering & Technology, Puttur, Karnataka, India putturpradeep@gmail.com
[2]Sri Siddhartha Institute of Technology, Tumakuru, Karnataka, India
[3]SDM Institute of Technology, Ujire, Karnataka, India

(e) Level 4 - proliferative DR

Figure 1. Stages of DR as annotated by renowned opthalmologist

The main challenge in Deep Neural Networks (DNN) is overfitting, even when clustering algorithms perform well. To tackle this, feature selection can be employed to extract relevant information from the dataset, aligning with the application's requirements. These selected features offer valuable insights, enabling deeper exploration of possibilities[5]. This research aims to enhance the synergy between deep learning methods and clustering algorithms, enabling the utilization of unattended acquisition techniques for efficient dimensionality reduction classification.

$$C = \{x : \| x - C \| 2 \leq \| x - \mu \| 2 \forall \square \} \quad \text{-----(1)}$$

$$\square\square = |\square\square| \sum \square\square \in \square\square \square\square \quad \text{- - - - - (2)}$$

This iterative process in the equations (1) and (2) minimizes the within-cluster sum of squares, leading to the optimal clustering of the data items.

The traditional k-means algorithm, a key method for vector quantization, has specific limitations that impact its effectiveness. Firstly, it operates linearly, meaning data elements assigned to one of the batch are not considered into account for others. However, in real- world situations, data items often share attributes across clusters[4]. Secondly, the algorithm tends to converge to local minima rather than the global minimum due to its dependence on the initial random selection of centroids, which can negatively impact the final cluster configuration. Lastly, the k-means algorithm makes minimal adjustments to cluster centers during each iteration, resulting in prolonged convergence times. Therefore, the proposed work integrates k-means clustering with local cluster information, calculated using the inverse Euclidean distance and cluster center.

The proposed algorithm is evaluated using fundus images for DR detection, with a comparative analysis conducted against the Updated Fuzzy C Means (UpFCM) algorithm. This analysis also includes Fuzzy C-Means (FCM), K-Means, and Autoencoder-based Deep Embedded Clustering (DEC). This comparison aims to validate the model's generalizability, and notably, the planned model outperformed the other algorithms.

The basic k-means grouping algorithm is enhanced by incorporating local information parameters for specific clusters. This enhancement involves calculating the inverse euclidean- distance betwixt data items and cluster centers, serving as an adjustment factor that expedites the convergence of cluster centers and reduces the amount of iterations required. The planned method is compared with the updated fuzzy f-means algorithm, which optimizes global optima within the conventional FCM framework. This revised algorithm includes an additional median adjustment parameter in the objective-inclusion method to further improve solution optimality.

2. LITERATURE REVIEW

Supervised learning frameworks hold momentous promise for solving various problems, but unsupervised learning methods can reveal new opportunities. Specifically, techniques for clustering in data mining can organize unfamiliar data into meaningful structures without explicit guidance, often using distance measurements via merging deep-learning together grouping results in deep clustering algorithms. DEC (Deep Embedded Clustering) is an effective method for unsupervised learning, excelling at clustering data. It assigns data items to clusters and learns useful attributes from the data, addressing gaps that supervised learning cannot fill. Innumerable methods have been enforced to enhance DEC. In DEC, Autoencoders create a feature space by transforming actual data into different features in a hidden space. The clustering process influences how the Autoencoder training is performed by setting clustering rules. DEC operates in two phases: first, a pre-training phase to set up initial parameters like cluster centers and stopping criteria; then, a fine-tuning phase where feature learning and clustering are performed together. DEC prefers Autoencoders due to their simplicity and reliability, and effective for reconstructing data[6][7].

However, the Discriminately-Boosted-Clustering(DBC) algorithm closely mirrors the procedures of DEC, considering the upbringing approach, clustering methodology, utilization of KL deviation, and the distribution of soft cluster assignments. The key difference is the application of CNN supported Autoencoder alternate to traditional Feedforward Autoencoder. This modification significantly improves the accuracy of the DEC technique, particularly when employed with image-based datasets. The Deep Clustering Network (DCN) approach addresses the challenge of integrating unattended learning into deep-learning by creating a unified operation. As an alternative of working with the more complex latent feature space, DCN utilizes the learned weights (typically denoted as "w") acquired after every training for the clustering task. This approach promotes a combined operation of dimensionality reduction and clustering, specifically employing k-means clustering assignment to achieve its objective[8][9]. The optimization criteria in this

process involve three practical steps: (1) reducing dimensionality, (2) reconstructing data, and (3) enhancing cluster-based regularization.

While previous algorithms like DEC, DBC, and DCN have addressed certain issues, they face limitations in scalability and efficiency, especially with large datasets. Deep Embedded Regularized Clustering (DEPICT) sets itself apart with a novel approach. Instead of conventional methods, DEPICT employs a multi-layer convolutional autoencoder and introduces a relative entropy-based objective function to regulate cluster assignments. To enhance robustness, DEPICT uses a data-dependent regularization strategy to compute the reconstruction loss, preventing overfitting during network training. This study presents a joint learning framework that efficiently minimizes both the clustering loss and the reconstruction loss while concurrently training the network[10][11][12]. Continuing the exploration of unsupervised deep-learning, experts have experimented a novel approach called VaDE, which utilizes the Variational Autoencoder (VAE). VaDE performs unsupervised generative clustering by combining the Gaussian-Mixture-Model(GMM) with a Deep-Neural- Network(DNN). This approach allows for interpolation within the latent representation space and the generation of new samples that align with the data distribution[13][14].

In VaDE, data clustering begins with the employment of the GMM, followed by generating latent embedding features denoted as "z." The properties are fed into the DNN for decrypting into observable data. The combined efforts of the VAE and GMM is governed by the Evidence-Lower-Bound (ELBO), as demonstrated in the research. Deep learning and clustering are two aspects of the Joint-Unsupervised- Learning (JULE) approach's data processing [15][16].

Using a stacked CNN, the method is particularly effective with image datasets. This approach uses a recurrent structure, where CNNbased learning takesplace on the backwa rd pass and agglomerative clustering is applied on the forward pass. Unlike other methods, JULE presents a single loss function that maximizes the recurrent architecture, which includes both CNN and agglomerative clustering elements. One important consideration while describing the drawbacks and benefits of clustering algorithms is the data's dimension. Grouping methods are typically exploited to calculate the gap between data items to evaluate similarities, but as data dimensionality rises in complexity, this method becomes less useful. Deep learning methods assist this, although dimension reduction is still required for spectral clustering techniques as well. The improvement of a clear verifiable utility for deep-learning framework training, can lead to an improvement in the demonstration of the deep clustering algorithm on high-dimensional data[17][18].

3. METHODOLOGY

Recently, some notable breakthroughs within the realm of automated disease identification, which have greatly benefited society and strained medical and healthcare systems. Several elements have played a role in this advancement, including superior image capture quality, the use of artificial intelligence, vast storage capabilities, and enhanced computing power[19][20]. In today's world, there's a rising need for data analysis and insights by organizations focused on growth, aiming to improve their results, efficiency, and performance. The significance of data analysis and learning is on the increase, leading to promising outcomes that offer efficient solutions for future improvements. Although traditional algorithms and their diversification have been utilized to improve results, there's a growing focus on developing more effective solutions. There's a need to update the way clustering/deep learning algorithms are processed due to their shortcomings[22]. The problem of over-fitting is a major issue in deep learning networks, even when the clustering algorithm performs well. However, by selecting relevant features for the application, it's feasible to eliminate the unnecessary matter from the data-set. Moreover, the selected features provide constituent subject matter that permits a deeper exploration of other opportunities.

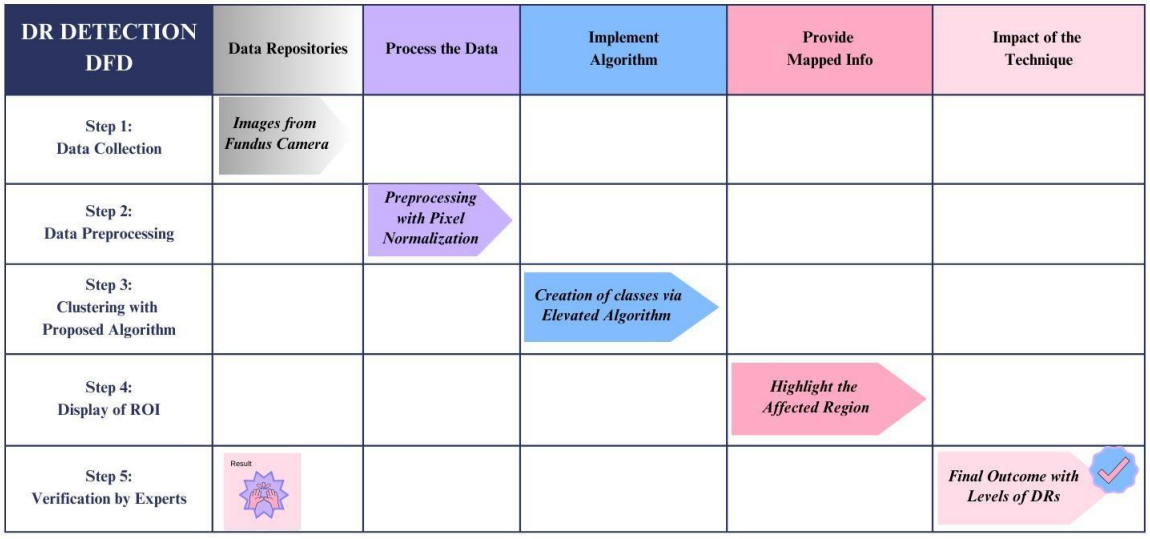


Figure 2. The flow of data via the incorporated technique

This method(represented in figure 2) develops a system for detecting disaster recovery that uses imprecise data together with the K-means (FIKM) method. Afterward, it's evaluated against the UpFCM algorithm to show its applicability across different scenarios. Throughout this method, statistical characteristics can be either created or identified to form a unique mix of distinct independent traits.

3.1 Data repository

Within the research framework, the datasets employed encompass IDRiD, DIATREB1, and fundus photographs - DIATREB2. These fundus photographs, were from numerous patients diagnosed with moderate to severe Diabetic Retinopathy (DR) affecting both eyes. These images were captured mostly with a Zeiss tabletop fundus camera, utilizing a 35mm focal length. The dataset's statistical overview is presented in table 1. The IDRiD dataset was developed from actual clinical eye examinations conducted at a clinic in Nanded, India, with a focus on individuals with diabetes[21]. The images captured have a 50-degree field of view, a resolution of 4288 × 2848 pixels, and are saved in jpg format, each with a 50-degree field of view, divided into five categories of Diabetic Retinopathy (DR) severity levels ranging from 0 to 4, as detailed in table. All the imageries in the data repository is annotated and verified by a senior ophthalmologist.

Table 1. Fundus repository in the experiment

Name of data-set	Quantity	Instrument	Form	Format	Comments
DIARET DB0	150	50°-FOV DFC	Fundus	PNG	DR discovery&marking
DIARET DB1	92	50°-FOV DFC	Fundus	PNG	DR discovery& marking
IDRID	516	NR	Fundus	JPEG	DR-grading and lesion separation

3.2 Preprocesing

Preprocesing is essential for maximizing the quality of images and enabling precise analysis. These photos frequently have a number of issues, including noise, the lighting is not consistent, and there are artifacts like dust spots or reflections. Preprocesing techniques such as contrast enhancement, noise reduction, and artifact removal are necessary to guarantee that important anatomical structures such as the macula, optic disc, and blood vessels can be clearly seen.
Normalization of pixel values contributes to consistency between images, and pre- processing-enabled segmentation methods are crucial for differentiating particular structures relevant to diagnosis of DR, like anomalies in blood vessels. Furthermore, by adding variances to the training dataset, data augmentation during pre-processing strengthens the robustness of machine learning models and ultimately increases the effectiveness of automated analysis and DR detection.

3.3 Implementation

The traditional k-means approach, a vector quantization method, is recognized as an effective clustering technique in data mining. It uses an iterative process to reach convergence, beginning with the random selection of cluster centers and adjusting distances. In each iteration, cluster-centers are recalculated by mean of data-items in each cluster from the former iteration. This procedure prolongs until an intersection, with the k-centers progressively moving closer to their final positions. The k-means technique is established in following ways. For a data point x_i and a cluster center μ_k the distance is calculated as:

$$d(x_i, \mu_k) = \sqrt{\sum_{j=1}^{n} (x_{ij} - \mu_{kj})^2}$$

where x_{ij} is the j-th feature of the i-th data item, and μ_{kj} is the j-th feature of the k-th cluster center. Each data item x_i is assigned to the batch whose center is closest. This is constructed by reducing the distance:

$$C_i = \min. d(x_i , \mu_k) \text{ - - - - - (4)}$$

where C_i is the cluster assignment for data item x_i.

After assigning all data items to clusters, the cluster centers are updated by calculating the mean of all data items assigned to each cluster. For cluster k, the new center μ_k is computed as:

where $|C_k|$ is the number of data items assigned to cluster k, and C_k is the amount of data items in cluster k.

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i - - - (5)$$

The formula iterates through the assignment and update steps until the batch centers no longer change significantly or a supreme amount of repetition is reached. Convergence can be checked by evaluating if the change in cluster centers is below a small threshold ϵ:

$$\|\mu_k^{(t+1)} - \mu_k^{(t)}\| < \epsilon - - - (6)$$

where $\mu^{(t)}_k$ and $\mu^{(t+1)}_k$ are the cluster centers at iterations t and t+1, respectively.
The Convergence function, denoted as Ci, measures the variation between cluster centers from the previous and current iterations. The stopping criterion and control of the iterative process depend on the precision level is specified. The variable n represents the number of clusters, either predefined or requested. In this context, the variables are provided. The current cluster center, denoted as μ(t), and the updated cluster center, denoted as μ(t+1), are calculated using the cluster assignment operation.
The amount of cluster bags available, denoted as Ci, corresponds to the user-specified number of clusters, as the bags are created to match this required quantity. The algorithm uses a selected set of data items, denoted as D, which represent the closest items or elements to the batch centers. These data items are crucial for deciding the cluster centers during the
clustering process. The variable |D|represents the cardinality of the dataset, indicating the total number of elements.

The traditional k-means algorithm has several limitations in handling data effectively. One such limitation is its linear processing approach, which confines data elements within a single cluster and does not allow them to influence other clusters. In real-world scenarios, there is often significant overlap. Additionally, the k-means tends to converge to localized minima instead of globalized minima, which means the final clusters can be heavily affected by the initial random cluster center placements. The minimal movement of data items towards cluster centers in each iteration affects the overall time complexity of the process. Nonetheless, by integrating local information through fuzzy clustering techniques, k-means can better capture complex cluster structures and accurately determine clusters when they converge or have irregular shapes.
FIKM combines the principles of fuzzy clustering with informative constraints to enhance the clustering process. This method incorporates both fuzzy membership and additional information to refine cluster assignments. In Fuzzy k-means, each data point xi has a inclusion value u_{ik} for each cluster k, representing the degree of belonging of xi to cluster k. The membership function satisfies:

$$\sum_{k=1}^{n} u_{ik} = 1 - - - (7)$$

where u_{ik} is the membership degree of data item x_i in cluster k, and n with number of clusters. The objective-method for fuzzy clustering aims to decrease the weighted total of squared distances between data-items and cluster centers, adjusted by the membership values. The objective function J is given by:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{n} u_{ik}^m \cdot \|x_i - \mu_k\|^2 - - - (8)$$

where:

- m is the fuzziness parameter (typically m>1).
- $x_i$is the i-th data point.
- μ_k is the center of cluster k.
- ‖.‖ denotes the euclidean distance. For cluster k, the updated center μ_k is:

$$\mu_k = \frac{\sum_{i=1}^{m} u_{ik}^m \cdot x_i}{\sum_{i=1}^{m} u_{ik}^m} - - - (9)$$

where $\sum \square\square=1\square$

□□□
is the amount of membership values raised for k

The rank values are updated according to gap between data-items and cluster centers. For each data item x_i and cluster k, the membership value u_{ik} is updated as:

$$u_{ik} = \frac{1}{\sum_{j=1}^{n} \left(\frac{\|x_i - \mu_k\|}{\|x_i - \mu_j\|} \right)^{\frac{2}{m-1}}} - - - (10)$$

where $\|x_i - u_k\|$ is the gap between the data items and center of batch u_k and [2/(m-1)] is the exponent in the denominator.

Fuzzy Informative k-means may also incorporate additional constraints or information to improve clustering. These constraints can be represented as additional terms in the objective function or as penalty terms that adjust the membership values based on external information.

The Fuzzy Informative k-means algorithm iteratively updates the membership values and cluster centers until convergence, often defined by a small change in the objective function or membership values between iterations. The equation represents the objective function, which evaluates the level of convergence achieved during the iterative process. It utilizes the current membership functions and the local information parameter to estimate the convergence state. Thus, the application of the fuzzy approach to the k-means algorithm is demonstrated through the systematic procedure outlined below.

The algorithm provides a more flexible clustering approach by allowing each data point to have partial membership in multiple clusters, which is useful for handling overlapping or ambiguous data. The preference for fuzzy informative k-means over other methods is due to its ability to handle noisy and high-dimensional data effectively, while also providing robustness against outliers. This resilience ensures that noisy data items do not unduly affect the clustering results. Additionally, fuzzy informative k-means combines the interpretability of traditional k-means with the flexibility of fuzzy clustering, making it well-suited for various real-world applications where data frequently exhibits complex patterns and noise.

3.4 Updated Fuzzy C-Means

Fuzzy C-Means(FCM) is an unsupervised-soft batch operation, that represents its clusters as fuzzy sets and thus is quite distinct from k-means as it allows overlapping clusters and can represent data items as elements of more than one cluster. Given a set of number of data items with cluster indicators, denoted as $\{x_i, c_i\}$, the objective function to be minimised in an iterative process is defined as: where u_{ij} represents the degree to which xi belongs to cluster j and m = 2. A termination condition requires convergence of the method, usually measured by reducing the mean squared error. FCM operates in the realm of unsupervised learning, and hence the clustering procedure initiated with random initialization. Fuzzy-C-Means(FCM), extends k-means in a straightforward way: it allows each data point to belong to more than one cluster with a different degree of participation. This approach is useful when the data items in a cluster form a fuzzy region as opposed to a hard division between neighbourhoods. Here's a step-by-step account of what FCM does.

1. In the beginning, define the number of clusters k to be identified. Choose the fuzziness parameter (usually (m > 1)), which controls the level of ambiguity in the clustering. Initialize the cluster centers randomly or using a heuristic method. Initialize the inclusion(membership) matrix (U), where (u_{ik}) represents the degree of inclusion of data point x_i in cluster k. Ensure that: $\sum_{□=} C□ = 1$ and $0 <= u_{ik} <= 1$
2. Compute the new cluster centers based on the membership value u_{ik}

$$\mu_k = \frac{\sum_{i=1}^{m} u_{ik}^m \cdot x_i}{\sum_{i=1}^{m} u_{ik}^m} - - - (11)$$

where x_i is the I-th data item, u_k - center of cluster k and m - the fuzziness parameter.

3. Update the membership values settled on the distances between data items and cluster centers.

$$u_{ik} = \frac{1}{\sum_{j=1}^{k} \left(\frac{\|x_i - \mu_k\|}{\|x_i - \mu_j\|} \right)^{\frac{2}{m-1}}} - - - (12)$$

where $\|x_i - u_k\|$ is the euclidian distance between the data item x_i and the cluster center u_k, and 2/(m-1) is the exponent

used in the calculation.

4. The objective function (J) will be calculated to measure the clustering performance

$$J = \sum_{i=1}^{m} \sum_{k=1}^{k} u_{ik}^{m} \cdot \|x_i - \mu_k\|^2 - - - (13)$$

where $\|x_i - u_k\|^2$ is the squared distance between the data item and the cluster center.

5. Evaluating convergence by checking, if the change in the target function or the membership(inclusion) matrix (U) falls below a pre-defined threshold.

$$\|U^{(t+1)} - U^{(t)}\| < \epsilon - - - (14)$$

where U^t and U^{t+1} are the membership matrices at iterations (t) and (t+1), respectively.

6. Iterate the steps 2 through 5 until convergence is achieved. This typically involves iterating until the change in cluster centers or the membership matrix is minimal, and record the events for the occurences.

While FCM often produces superior results, it is limited by its susceptibility to local optima, which can increase the time complexity required to reach convergence. In certain situations, convergence might not be reached at all. To address this issue, we have introduced a median adjustment parameter to improve performance. This parameter introduces a penalty term that guides cluster centroids towards the cluster's median rather than relying solely on the data items. By directing the centroids based on the median rather than the mean, this adjustment enhances the algorithm's robustness, stability, and convergence properties. The movements also recorded via a data-structure. This provides a flexible approach to clustering, accommodating data with overlapping clusters, involving noisy or sparse data and offering insights into the degree of membership for each data item in multiple clusters.

The below factors are taken into account for running the algorithm.

1. Initialize all the parameters like, desired number of clusters, the fuzzy factor, convergence criteria/status maximum allowable iterations.
2. The data-vector vector is subjected to fuzzification and afterwards reorganized according to a random membership function.
3. The dimension is utilized to generate random memberships, and its variants exhibit a soft clustering nature, implying the mutual association of data items in two or more clusters. Here, representing the absolute amount of data items available in, and, signifying the user- determined number of clusters, is utilized.
4. The membership function is premeditated at each iteration using the latest cluster centre. If the convergence state is not achieved on this calculation, a new is computed.
5. The proposed variant of the FCM method distinguishes itself from another iterations by employing a distinct method for calculating its cluster centre in contrast to wherein the membership function is a crucial factor. The idea is to minimize iterations, thereby achieving the shortest possible convergence time.
6. The distance between the prior cluster centre and the present cluster is computed to promote better parameter selection and faster convergence.
7. While generating new updated cluster,the above procedure is repeated until the convergence state is attained.

4. RESULTS AND DISCUSSION

We mainly concentrated on enhancing existing clustering algorithms to identify an optimal solution. According to the results, the proposed algorithm effectively extracted maximum information from fundus images. The assessment is based on approximately 758 fundus images from various databases. The evaluation is performed on the original fundus images. The input is supplied in batches and found a very promising outcome, all the results are verified by the expert ophthalmologist. The below table 2 gives an analysis over other previous findings under clear situation in the data-repository.

Table 2. Analysis of clustering algorithms for accuracy(noiseless)

Epochs	k-means	FIKM	FCM	FLICM	EnFCM	FGFCM	MaFCM	UpFCM
Ep1	83.86	90.03	80.52	85.57	93.69	87.62	96.57	96.0
Ep2	82.38	89.40	79.53	85.68	90.23	87.38	97.97	96.87

Ep3	83.26	89.63	79.42	84.19	81.11	86.60	95.02	95.46
Ep4	84.28	91.55	80.63	84.60	83.38	89.40	97.41	96.68
Ep5	81.57	90.23	79.26	83.21	82.97	86.77	96.03	96.77
Ep6	82.40	88.70	79.02	83.27	81.32	88.02	98.34	97.78
Ep7	82.50	89.03	78.63	85.10	85.20	86.52	99.03	98.99
Ep8	82.73	88.90	77.63	83.64	79.27	85.14	95.99	96.97
Ep9	83.91	91.51	81.34	86.96	79.30	87.54	96.73	97.76
Ep10	82.54	88.12	79.45	84.27	82.63	85.61	98.14	97.89
Ep11	81.26	90.65	80.26	84.78	81.53	87.67	96.32	98.97
Ep12	82.28	89.72	81.02	85.29	83.54	88.02	98.47	98.98
Ep13	81.76	90.03	78.83	83.60	81.98	86.62	97.03	99.01
Ep14	82.24	88.55	79.23	84.21	82.23	86.14	98.64	99.13
Ep15	83.65	91.87	81.34	85.61	86.20	87.54	98.03	99.42
Ep16	83.67	89.11	79.43	85.76	84.84	86.21	98.11	99.39
Ep17	83.66	89.23	79.56	86.12	84.79	86.32	98.17	99.41
Ep18	83.69	87.45	79.79	85.91	85.56	87.46	98.21	99.43

The algorithm, which utilizes a fuzzy median strategy recordings, distinguishes itself by deriving median values from neighboring pixels, thereby enhancing accuracy on average. Table III provides a analysis of various techniques based on their accuracy under noisy atmosphere. Managing pixel contamination is challenging in complex data categorization or clustering, as noise can appear at various frequencies. The table shows batch-wise comparison among diversified techniques.

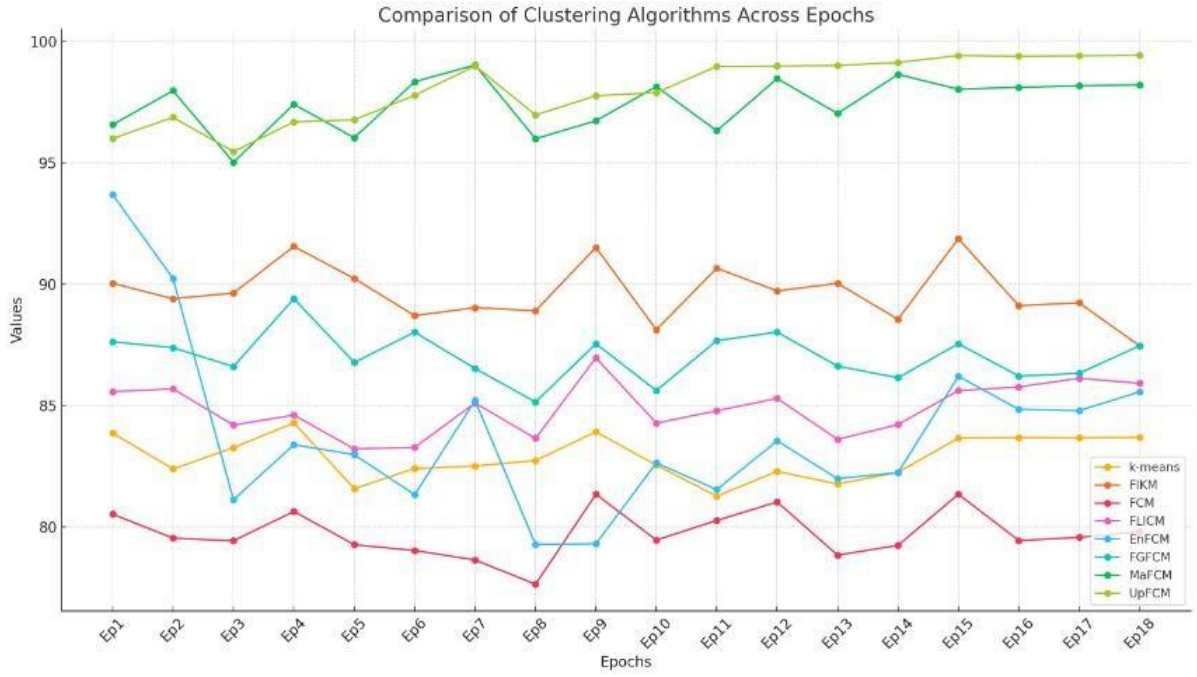


Figure 3 Performance of different clustering algorithms across the given epochs

TABLE 3. Analysis of algorithms(Noisy)

Batches	k-means	FIKM	FCM	EnFCM	FLICM	FGFCM	MaFCM	UpFCM
B1	91.65	96.64	92.05	92.29	94.13	87.62	97.63	96.77
B2	85.56	92.45	86.53	86.82	89.12	87.64	93.4	97.78
B3	77.13	87.70	81.28	81.31	85.93	82.52	90.19	98.99
B4	62.86	81.76	67.73	67.82	72.37	67.87	83.62	98.34
B5	63.48	82.11	80.08	76.67	73.67	71.83	83.12	97.59

Deep learning algorithms often depend on GPUs or high-performance computing facilities to accelerate the learning process. In this context, it is crucial that, techniques utilized for data processing operates within reasonable time constraints. In this study we have implemented with Intel i10 tenth-generation processor with a 2GB GPU and 16GB RAM considered for the computational resources available for efficient deep learning algorithm simulations.
Table 4 offers an overview of the average time taken by the algorithm during the simulation. Clustering algorithms typically make key decisions, such as selecting initial points, during the first iteration. Therefore, it is important that the processing time matches the system's capabilities to guarantee that the overall learning process remains within acceptable time limits.

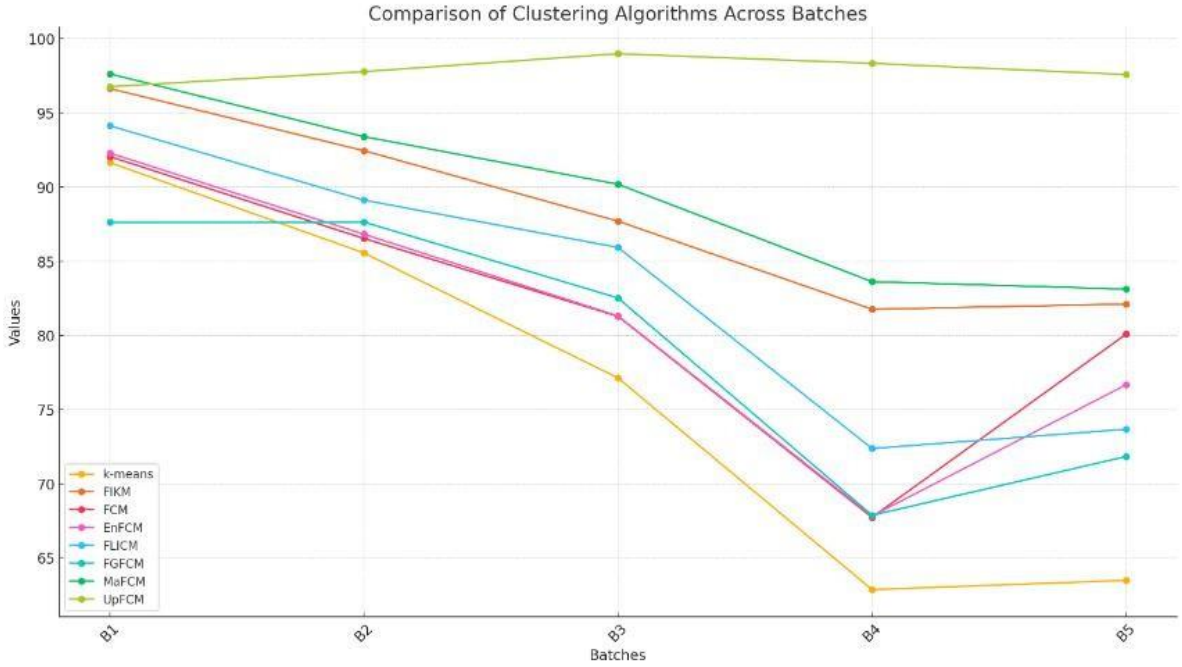


Figure 4. Analysis of techniques under noisy conditions

Table 4. Average time taken for executions

Execution no.	k-means	FIKM	FCM	EnFCM	FLICM	FGFCM	MaFCM	UpFCM
1	15.8	17.32	21.45	20.69	17.52	17.67	17.11	16.01
2	15.34	18.45	21.11	21.13	17.18	17.33	17.92	16.99
3	16.81	18.12	20.81	20.54	18.23	18.20	18.02	16.78
4	15.84	17.43	21.23	20.92	18.60	18.11	16.93	15.81
5	15.87	18.20	20.89	20.28	17.92	18.43	17.86	15.78
6	15.76	18.41	20.31	20.92	17.67	18.46	17.76	15.76
7	15.52	18.43	20.35	20.36	17.57	18.47	17.45	15.67

8	16.11	18.47	20.36	20.97	17.49	18.41	17.38	14.98
9	15.79	18.52	20.12	20.40	17.41	18.23	17.88	14.87
10	15.88	18.54	20.49	20.19	17.39	18.48	17.81	14.73

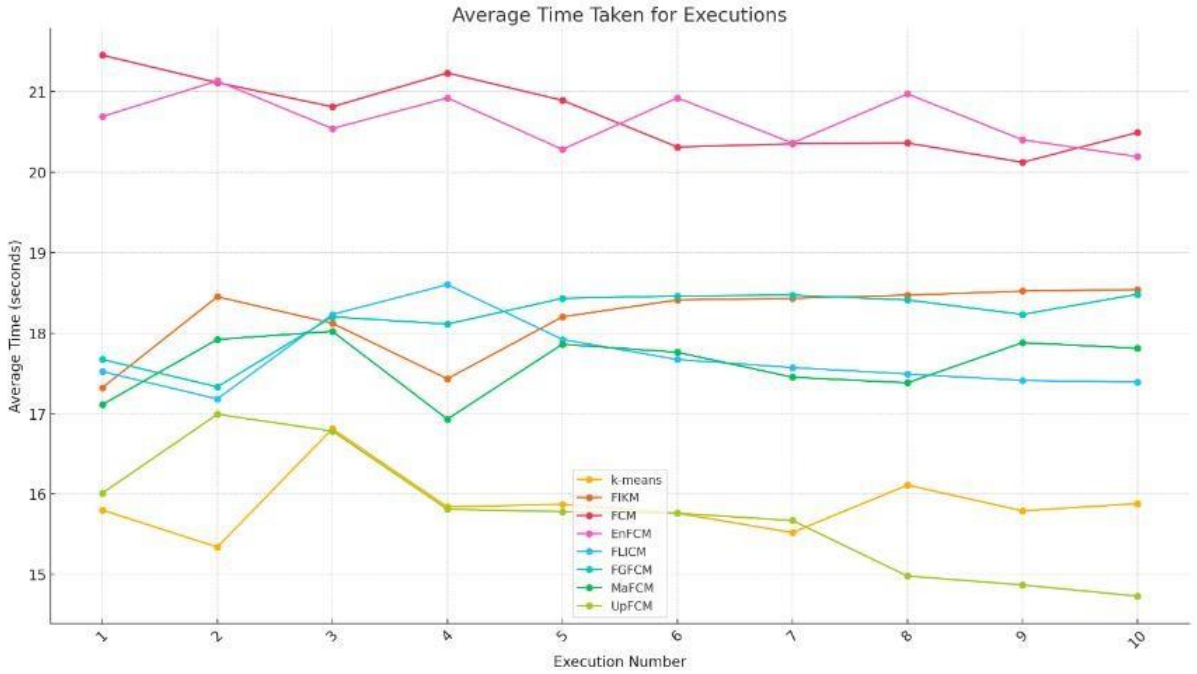


Figure 5. Analysis of computational complexities

The main goal of this study is to enhance the efficiency of unsupervised learning methods. Several challenges are addressed, including feature correction due to clustering loss insecurity, model complexity leading to implementation and training difficulties, hyper- parameter tuning, and the presence of overlapping clusters.This approach aims to establish a robust evaluation framework to achieve definitive results in enhancing unsupervised learning performance.
With unsupervised learning, without any labeled datasets where only the inputs are available, the goal is to identify hidden patterns or structures. The clustering accuracy, the goal is to group similar data items and to evaluate the degree to which these clusters align with known categories or segments. Normalized Mutual Information (NMI) is used to assess the clustering accuracy, these metrics compare the clustering results with the known ground truth to determine how well the clusters align. These metrics help in assessing how well-separated and cohesive are the clusters. Lower reconstruction error indicates that the dimensionality reduction has preserved the data's important features. Anomaly Detection Accuracy is used to identify rare or unusual data items. Accuracy here can be tricky to define, but it often involves evaluating metrics like precision, recall, or the F1 score. The evaluation often involves using metrics that measure the quality of patterns or structures identified by thealgorithm, to some known data (when available) or based on internal characteristics of the data.

$$\text{Accu} = \frac{\text{Num\_of\_correct\_clustered\_items}}{\text{Total\_Num\_of\_Items}} \quad \text{- - - - - (15)}$$

$$I(C;L) = \sum_{c \in C} \sum_{l \in L} p(c,l) \log \frac{p(c,l)}{p(c)p(l)} \quad - - - (16)$$

The Normalised Mutual Information (NMI) compares the clustering results with a ground truth. Mutual Information (MI) is the quantity of information gained regarding a single variable from another. NMI adjusts MI to account for the sizes of clusters or categories, outputting a metric from 0 to 1. A higher NMI indicates a better alignment of clusters with true labels. Now calculate the mutual information between the cluster assignments and the true labels. Call H(C) the entropy of the cluster assignments, and H(L) the entropic state of the true labels. Calculate I(C; L), the mutual information between (C) (clusters) and (L) (true labels). This is given by:
where p(c, l) is the joint probability of a point being in cluster (c) and having true label (l), and p(c) and p(l) are the marginal probabilities.
Now, Compute the entropy of clusters H(C) and the entropic state of true labels H(L)

$$I(C;L) = \sum_{c \in C} \sum_{l \in L} p(c,l) \log \frac{p(c,l)}{p(c)p(l)} - - - (17)$$

$$H(L) = -\sum_{l \in L} p(l) \log p(l) - - - (18)$$

Normalize the MI score by the average entropic state of the clustering and true labels to get NMI

$$NMI(C,L) = \frac{I(C;L)}{\sqrt{H(C) \cdot H(L)}} - - - (19)$$

The above formula ensures that NMI ranges from 0 to 1, with 1 indicating a perfect match between clusters and true labels, and 0 indicating no mutual information. The interpretation is
NMI = 1: Informs perfect clustering where clusters perfectly align with the true labels.
NMI = 0: Informs No mutual information between the clusters and the true labels, indicating poor alignment.
0 < NMI < 1: This is the Partial alignment; the closer to 1, the better the clustering matches the true labels.
Unsupervised learning, on the other hand, aims at finding hidden structures in data. Adjusted Rand Index (ARI) Measures the agreement between the clustering results and the true labels, taking into account the possibility of groups being formed just by chance.
ARI= (RI −Expected RI)/(MaximumRI−Expected RI) - - - - - (20)

Let R denote 0 if two data items are in the same cluster in both clustering and 1 if they are in different clusters, and let be the total number of pairs of data items. Then the Rand Index (RI) is outlined, expected Rand Index (Expected RI) is RI - which is adjusted for chance, and MaximumRI is the maximum possible Rand Index. First, draw a contingency table. Each cell (i, j) represents the number of data items. Each point is in cluster i in one clustering and in cluster j in the other clustering. Then the formula below to measure RI is -
RI=(m+n)/(m+n+o+p) - - - - - (21)
Assume m with total amount of pairs of data items that are in the same cluster in both clustering. n with amount of pairs of data items that are in different clusters in both clustering. O is the number of pairs of data items that are in the same cluster in one clustering but different clusters in the other. p with amount of pairs of data items that are in different clusters in one clustering, but other batches with same clusters. The procedure works well because it effectively encodes even small differences in the neighbourhood of the data to make clustering easier, and also because the algorithm is fairly resistant to segmentation problems, and is especially geared to global optimisation; their value become more obvious over time.
During the experiments with the proposed algorithms, several outcomes and potential limitations were identified. The algorithms generally bring forth batches those more adaptable to complex data structures, particularly in noisy or high-dimensional contexts. However, challenges include determining the optimal number of clusters and selecting suitable fuzzification parameters, which can greatly influence the clustering results. The performance of the algorithms may also be sensitive to the initialization of cluster centers and the choice of distance metric, necessitating careful tuning for the best results. Additionally, while the algorithms are robust against outliers, extremely skewed distributions or highly overlapping clusters may still present difficulties. Overcoming these limitations often requires iterative experimentation, parameter tuning, and thorough validation to ensure the algorithms perform effectively across various datasets and scenarios.

5. CONCLUSION

Diabetic Retinopathy (DR) is a severe chronic condition impacting around one-third of diabetic patients globally. To tackle this widespread issue, developing an automated DR detection system using retinal fundus images is essential. However, pointing out the data and its meaning for supervised learning in medical imaging can be time-intensive. Therefore, unsupervised clustering methods provide a viable alternative by uncovering hidden patterns and relationships within the data. This study introduces an enhanced clustering algorithm that integrates a fuzzy local parameter into the K-Means algorithm to improve performance and achieve global optimization. The proposed algorithm achieves an accuracy rate of 95.9% and an average execution time of 17.09 seconds. When evaluated on various image samples under both noisy and noiseless conditions, the algorithm shows superior performance in noisy scenarios, reaching an accuracy of 96%.

6. FUTURE DEVELOPMENT NEEDS

The consideration of processing new techniques that can process retinal images in real-time to provide diagnosis and

intervention in near-time, which would be especially relevant in the context of telemedicine and remote healthcare environments. Making automated DR screening tools available in primary care settings can thus open up access to timely diagnosis and treatment for patients in resource-limited and low- and middle-income settings, particularly in underserved communities which lack access to specialised eye care providers. It is also important to ensure responsible and ethical use of AI-based screening tools for DR detection such that concerns relating to algorithmic fairness, bias, privacy and transparency are addressed to improve trust in algorithmic-based healthcare delivery. There is an opportunity to make use of mobile technologies (such as smartphone-based retinal imaging devices and mobile apps) for cost-effective deployment of scalable DR screening and monitoring, especially in resource-limited settings, for early intervention and timely treatment of DR.

REFERENCES

1. Shoaib, M. R., Emara, H. M., Zhao, J., El-Shafai, W., Soliman, N. F., Mubarak, A. S., & Esmaiel, H. (2024). Deep learning innovations in diagnosing diabetic retinopathy: The potential of transfer learning and the DiaCNN model. Computers in Biology and Medicine, 169, 107834.
2. Joshi and P. T. Karule, "Detection of hard exudates based on morphological feature extraction," Biomed. Pharmacol. J., vol. 11, no. 1, pp. 215–225, 2018, doi: 10.13005/bpj/1366.
3. Pradeep Kumar KG, Dr. Karunakara K, Dr. Thyagaraju GS, Dr. Sunanda Dixit, A Data Mining Inspired Methodology towards the Identification of Diabetic Retinopathy, 1st International Conference on Computational Research and Data Analytics(ICCRDA 2020), IOP Conference Series: Materials Science and Engineering, Volume 1022, 012082
4. D. Nagpal, S. N. Panda, M. Malarvel, P. A. Pattanaik, and M. Zubair Khan, "A review of diabetic retinopathy: Datasets, approaches, evaluation metrics and future trends," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 9, pp. 7138–7152, 2021, doi: 10.1016/j.jksuci.2021.06.006.
5. J. Enguehard, P. O'Halloran, and A. Gholipour, "Semi-supervised learning with deep embedded clustering for image classification and segmentation," IEEE Access, vol. 7, pp. 11093–11104, 2019, doi: 10.1109/ACCESS.2019.2891970.
6. A. Kumar, L. Bi, J. Kim, and D. D. Feng, Machine learning in medical imaging. Elsevier Inc., 2019.
7. N. Hatipoglu and G. Bilgin, "Cell segmentation in histopathological images with deep learning algorithms by utilizing spatial relationships," Med. Biol. Eng. Comput., vol. 55, no. 10, pp. 1829–1848, 2017, doi: 10.1007/s11517-017-1630-1.
8. V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," Inf. Sci. (Ny)., vol. 250, pp. 113–141, 2013, doi: 10.1016/j.ins.2013.07.007.
9. S. Fogel, H. Averbuch-Elor, D. Cohen-Or, and J. Goldberger, "Clustering-Driven Deep Embedding With Pairwise Constraints," IEEE Comput. Graph. Appl., vol. 39, no. 4, pp. 16–27, 2019, doi: 10.1109/MCG.2018.2881524.
10. M. Z. Atwany, A. H. Sahyoun, and M. Yaqub, "Deep Learning Techniques for Diabetic Retinopathy Classification: A Survey," IEEE Access, vol. 10, pp. 28642–28655, 2022, doi: 10.1109/ACCESS.2022.3157632.
11. H. Naz, R. Nijhawan, and N. J. Ahuja, "An automated unsupervised deep learning–based approach for diabetic retinopathy detection," Med. Biol. Eng. Comput., vol. 60, no. 12, pp. 3635–3654, 2022, doi: 10.1007/s11517-022-02688-9.
12. M. Kampffmeyer, S. Løkse, F. M. Bianchi, L. Livi, A. B. Salberg, and R. Jenssen, "Deep divergence-based approach to clustering," Neural Networks, vol. 113, pp. 91–101, 2019, doi: 10.1016/j.neunet.2019.01.015.
13. T. Saba, S. T. F. Bokhari, M. Sharif, M. Yasmin, and M. Raza, "Fundus image classification methods for the detection of glaucoma: A review," Microsc. Res. Tech., vol. 81, no. 10, pp. 1105–1121, 2018, doi: 10.1002/jemt.23094.
14. S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, "Current techniques for diabetes prediction: Review and case study," Appl. Sci., vol. 9, no. 21, 2019, doi: 10.3390/app9214604.
15. J. Ganesan, A. T. Azar, S. Alsenan, N. A. Kamal, B. Qureshi, and A. E. Hassanien, "Deep Learning Reader for Visually Impaired," Electron., vol. 11, no. 20, pp. 1–22, 2022, doi: 10.3390/electronics11203335.
16. Y. Ren, K. Hu, X. Dai, L. Pan, S. C. H. Hoi, and Z. Xu, "Semi-supervised deep embedded clustering," Neurocomputing, vol. 325, pp. 121–130, 2019, doi: 10.1016/j.neucom.2018.10.016.
17. W. Wang, D. Yang, F. Chen, Y. Pang, S. Huang, and Y. Ge, "Clustering with Orthogonal AutoEncoder," IEEE Access, vol. 7, pp. 62421–62432, 2019, doi: 10.1109/ACCESS.2019.2916030.
18. T. Melo, A. M. Mendonça, and A. Campilho, "Microaneurysm detection in color eye fundus images for diabetic retinopathy screening," Comput. Biol. Med., vol. 126, no. September, 2020, doi: 10.1016/j.compbiomed.2020.103995.
19. T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy c-Means algorithms for very large data," IEEE Trans. Fuzzy Syst., vol. 20, no. 6, pp. 1130–1146, 2012, doi: 10.1109/TFUZZ.2012.2201485.
20. D. C. Hoang, R. Kumar, and S. K. Panda, "Realisation of a cluster-based protocol using fuzzy C-means algorithm for wireless sensor networks," IET Wirel. Sens. Syst., vol. 3, no. 3, pp. 163–171, 2013, doi: 10.1049/iet-wss.2012.0132.
21. P. Porwal et al., "Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research," Data, vol. 3, no. 3, pp. 1–8, 2018, doi: 10.3390/data3030025.
22. Huma Naz, Tanzila Saba, Faten S. Alamri, Ahmed S. Almasoud, and Amjad Rehman, " An Improved Robust Fuzzy Local Information K-Means Clustering Algorithm for Diabetic Retinopathy Detection", doi: 10.1109/access.2017.