ANALYSIS OF CLUSTERING ALGORITHMS FOR COVID-19 IN CT IMAGES

Jyothi prasad M1, Arul P2, Sujitha S3, Swaathy B4, Swetha V P5
1,2 Assistant Professor, 3,4,5 U.G. Students,
Department of Electronics and Communication Engineering,
R.M.D Engineering College, Chennai, India
1 mjp.ece@rmd.ac.in 2 spa.ece@rmd.ac.in 3 sundararajansujitha@gmail.com 4 bswaathy23@gmail.com 5 swethaofficial31@gmail.com

ABSTRACT

Early detection and diagnosis are critical factors to control the COVID-19 spreading. Several machine learning-based methodologies have been recently proposed for COVID-19 screening in CT scans as a tool to automate and help with the diagnosis. In this project, we analyzed the various clustering algorithms like K means, improved k means and fuzzy c means. This method K-Means and Fuzzy logic have been proposed in order to recognize overlapping pattern classes (or) degraded images. The main objective of the project is to analyze the covid-19 abnormalities in CT lung images. The project involves extracting the abnormal regions and to spot the exact location of covid-19 in the CT lung image.

In K-Means the Process time is faster than other clusterings with more number of data points. It is very suitable to spot the abnormal region because of updating centroids in some cases with improved K-Means[1]. FCM gives better results for overlapped data. Image segmentation may be defined as a technique, which partitions a given image into a finite number of non-overlapping regions concerning some characteristics, such as gray value distribution, texture distribution, etc. segmentation of medical images is required for many medical diagnoses like radiation treatment planning volume visualization of regions of interest (ROI) defining the boundary of covid-19, etc. Algorithms in this project have been used for various applications such as Biomedical applications, Information Retrieval, Pattern Recognition, Object detection.

Keywords— Clustering algorithms, K means, Improved K means, Fuzzy C means, CT images, COVID-19.

I. INTRODUCTION

With the improvement in imaging technology, diagnostic imaging has become an essential tool in medicine nowadays. X-ray angiography (XRA), magnetic resonance angiography (MRA), magnetic resonance imaging (MRI), computed tomography (CT), and other imaging modalities are widely used in medical practice. Complimentary information about the patients is provided by such images. While increased size and volume in clinical images required the automation of the diagnosis process, the new advances in computer technology and fewer costs have made it feasible to develop such systems. Covid on clinical images forms an important step in finding the solution for many practical applications such as diagnosis of the CT lung diseases and registration of patient CT images obtained at different times. Segmentation methods provide the essence of medical image applications such as radiological diagnostic systems, multimodal image registrations, creating anatomical atlases, visualizations, and computer-aided surgeries. Covid segmentation algorithms are the important components of automated-radiological diagnostic systems. Segmentation methods may differ depending on the imaging modality, application domain, method being automatic or semi-automatic, and other specific factors. No segmentation method can extract vasculature from each medical image modality. While some methods develop pure intensity-based pattern recognition techniques such as thresholding followed by connected component analysis, some other methods use explicit Covid19 models to obtain the abnormal spot contours[3][5]. Depending on the image quality and the general image artifacts such as noise, some segmentation methods need image...
preprocessing before the segmentation algorithm. And also, some methods apply post-processing to solve the problems coming from over-segmentation[6].

Medical image segmentation algorithms and techniques can be classified into six important classes, pattern recognition technique, model-based approach, tracking-based approach, artificial intelligence-based approach, neural network-based approaches, and miscellaneous tube-like object detection approach. Pattern recognition techniques are further classified into seven classes, multi-scale approach, skeleton-based approach, region growing approaches, ridge-based approach, differential geometry-based approach, matching filters approach, and mathematical morphology schemes[6][7].

Clustering analysis plays an essential role in scientific research and commercial application[8]. This article gives a report of the present covid-19 segmentation techniques using clustering approaches and gives both old and recent literature corresponding to abnormal CT lung segmentation algorithm and techniques[4].

## II. IMAGE SEGMENTATION

Segmentation issues are the bottleneck to accomplish object extraction, object-explicit estimations, and fast object rendering from multi-dimensional picture information. Straightforward division strategies depend on nearby pixel-neighborhood characterization. Such strategies flop anyway to "see" worldwide items instead of neighborhood appearances and require frequently concentrated administrator help. The explanation is that the "rationale" of an item doesn't follow that of its neighborhood picture portrayal. Nearby properties, like surfaces, edges, and so on don't generally address associated highlights of a given item.

**Clustering:** Clustering can be viewed as the main unsupervised learning problem, in this way, it manages to discover a design in an assortment of unlabeled information. A cluster is in this manner an assortment of items that are "comparable” among them and are "disparate” to the articles having a place with different groups.

Clustering algorithms may be classified as listed below

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering

In the primary case, information is gathered selectively, so that assuming a specific datum has a place with a distinct group, it couldn't be remembered for another cluster. The subsequent kind, covering grouping, utilizes fluffy sets to cluster information, so that each point may have a place with at least two clusters with various levels of participation. For this situation, information will berelated to proper enrollment esteem. A progressive grouping calculation depends on the joining between the two closest clusters. The starting condition is acknowledged by setting each datum as a group. After a couple of emphases, it arrives at the last clusters wanted[4].

K-means is one of the least difficult unaided learning calculations that take care of the notable clustering issue. The method follows a basic and simple approach to group a given informational index through a specific number of clusters (expect k groups) fixed deduced. The primary thought is to characterize k centroids, one for each group. These centroids ought to be put in a guile route due to various areas that cause the distinctive outcome. In this way, the better decision is to put them however much as could be expected far away from one another. The subsequent stage is to take each guide having a place toward a given informational index and partner it to the closest centroid. At the point where no information is forthcoming, the initial step is finished and an early groupage is finished. Now we need to re-compute k new centroids as twofold focuses of the clusters coming about because of the past advance. After we have these k new centroids, another limiting must be done between a similar informational index focuses and the closest new centroid. A circle has been created. Because of this circle, we may see that the k centroids change their area bit by bit until no more changes are finished. As such centroids don't move anymore. At long last, this calculation targets limiting a goal work, for this situation a squared error function[11][12].

www.turkjphysiotherrehabil.org
TYPES OF IMAGES

Each pixel of an image is typically associated to a specific ‘position’ in some 2D region, and has a value consisting of one or more quantities (samples) related to the position. Digital images can be classified according to the number and nature of those samples:

- Binary (bi-level)
- Intensity (Gray Scale)
- RGB
- Indexed (False-color)

The term digital image is also applied to data associated to points scattered over a three dimensional region, such as produced by tomographic equipment. In that case, each datum is called a voxel[9].

III. REASON FOR CHOOSING CT

This Today’s medical image analysis algorithms are commonly divided into different classes Enhancement algorithms are used to reduce image noise and increase the contrast of structure of interest. Segmentation algorithms are used to delineate structures of interest in an image discriminate them from each other, and from background tissues. Quantification algorithms are used to extract essential diagnostic information such as shape, size, texture, angle, and motion. Registration algorithms are used to align two different images of the same part of the body. Visualization algorithms are used to facilitate visual inspection of, possibly multimodal medical and biological data. Finally, Compression and communication, algorithms are used to store, retrieve, and transfer medical images and associated patient information ineffective and convenient manner[13][14].

The body is largely composed of water molecules each water molecule has two hydrogen nuclei or protons. When a person goes inside the powerful magnetic field of the scanner, the magnetic moments of some of these protons change and align with the direction of the field. A radio frequency transmitter is briefly turned on, producing a further varying electromagnetic field. The photons of this field have just the right energy, known as the resonance frequency, to be absorbed and flip the spin of the aligned protons in the body. The frequency at which the protons resonate depends on the strength of the applied magnetic field. After the field is turned off, those protons which absorbed energy revert back to the original lower-energy spin-down state. They release the difference in energy as a photon, and the released photons are detected by the scanner as an electromagnetic signal, similar to radio waves.

Contrast agents may also be directly injected into a joint in the case of arthrograms, MRI images of joints. Unlike CT, MRI used no ionizing radiation and is generally a very safe procedure. Nonetheless the strong magnetic fields and radio pulses can affect metal implants, including cochlear implants and cardiac pacemakers. In the case of cochlear implants, the US FDA has approved some implants of MRI compatibility. In the case of cardiac pacemakers, the results can sometimes be lethal, so patients with such implants are generally not eligible of MRI[11].

IV. PARAMETERS USED

Entropy: In Image, Entropy is defined as corresponding states of intensity level which individual pixels can adapt. It is used in the quantitative analysis and evaluation image details, the entropy value is used as it provides better comparison of the image details. The higher value of Entropy = more detailed information. Entropy is a measure of image information content, which is interpreted as the average uncertainty of information source.

Specificity, Sensitivity and Accuracy:

There are several terms that are commonly used along with the description of sensitivity, specificity and accuracy. They are true positive (TP), true negative (TN), false negative (FN), and false positive (FP). If a disease is proven present in a patient, the given diagnostic test also indicates the presence of disease, the result of the diagnostic test is considered true positive. Similarly, if a disease is proven absent in a patient, the diagnostic test suggests the disease is absent as well, the test result is true negative (TN). Both true positive and true negative
suggest a consistent result between the diagnostic test and the proven condition (also called standard of truth). However, no medical test is perfect. If the diagnostic test indicates the presence of disease in a patient who actually has no such disease, the test result is false positive (FP). Similarly, if the result of the diagnosis test suggests that the disease is absent for a patient with disease for sure, the test result is false negative (FN). Both false positive and false negative indicate that the test results are opposite to the actual condition.

Sensitivity, specificity and accuracy are described in terms of TP, TN, FN and FP.

- Sensitivity = \( \frac{TP}{TP + FN} = \frac{\text{Number of true positive assessment}}{\text{Number of all positive assessment}} \)
- Specificity = \( \frac{TN}{TN + FP} = \frac{\text{Number of true negative assessment}}{\text{Number of all negative assessment}} \)
- Accuracy = \( \frac{TN + TP}{TN + TP + FN + FP} = \frac{\text{Number of correct assessments}}{\text{Number of all assessments}} \)

As suggested by above equations,

Sensitivity is the proportion of true positives that are correctly identified by a diagnostic test. It shows how good the test is at detecting a disease. Specificity is the proportion of the true negatives correctly identified by a diagnostic test. It suggests how good the test is at identifying normal (negative) condition. Accuracy is the proportion of true results, either true positive or true negative, in a population. It measures the degree of veracity of a diagnostic test on a condition.

V. PROPOSED METHOD

Clustering analysis, a significant innovation in information mining, is a viable technique for investigating and finding valuable data from various information. Cluster algorithm bunches the information into classes or groups so that objects inside a group have high comparability in contrast with each other, however are not at all like items in different bunches. Dissimilarities are surveyed dependent on the trait esteems portraying the items. Frequently, distance measures are utilized. In this venture, the input picture is changed over into 1D Image and different grouping strategies are applied. At long last fragmented outcome is acquired after 2D transformation and execution examination has made.

Fig. 1 Proposed Block diagram.

A. K-MEANS CLUSTERING

As a branch of statistics and an example of unsupervised learning, clustering provides us an exact analysis tool from the mathematic view K-means algorithm belongs to a popular partition method in cluster analysis. The most widely used clustering error criterion is squared-error criterion, it can be defined as

\[
J_c = \sum_{j=1}^{c} \sum_{k=1}^{n_j} \left\| x_{k}^{(j)} - m_j \right\|^2
\]

where \( J_c \) is the sum of square-error for all objects in the database, \( x_{k} \) is the point in space representing a given object, and \( m_j \) is the mean of cluster \( C_j \) [12]. Embracing the squared-error standard, K-means functions admirably when the groups are smaller mists that are fairly all-around isolated from each other and are not reasonable for finding clusters with no curved shapes or clusters of altogether different size. For endeavoring to
limit the square-error criterion, it will isolate the articles in a single bunch into at least two groups. Furthermore, while applying this square-error rule to assess the grouping results, the ideal clusters relate to the extremum. Since the target work has numerous nearby negligible qualities, if the aftereffects of instatement are by and large close to the neighborhood insignificant point, the calculation will end at a nearby ideal. Along these lines, irregular choosing starting bunch community is not difficult to get in the neighborhood ideal, not the whole ideal. For beating that square-error criterion is difficult to recognize the huge contrast among the bunches, one strategy has been created which depends on the representative point-based procedure.

Besides, there are various approaches to solving the problem that the performance of algorithm heavily depends on the initial starting conditions: the simplest one is repetition with different random selections. Some algorithms also employ simulation anneal technique to avoid getting into local optimal. The idea is that multiple sub-samples are drawn from the dataset clustered independently, and then these solutions are clustered again respectively, the refined initial center is then chosen as the solution having minimal distortion over all solutions. Aiming at the dependency to initial conditions and the limitation of K-means algorithm that applies the square-error criterion to measure the quality of clustering, this paper presents a new improved K-means algorithm that is based on effective techniques of multi-sampling and once-clustering to search the optimal initial values of cluster centers. Our experimental results demonstrate the new algorithm can obtain better stability and excel the original K-means in clustering results.

Fig. 2 Flow chart for K-means Clustering

Algorithm:

STEP 1: Read the input image
STEP 2: Get the number of Clusters
STEP 3: Calculate the centroid for each cluster
STEP 4: Calculate distance between each pixel to centroids
STEP 5: Group pixels based on minimum distance
STEP 6: Check No of pixels to move centroid IF Yes End
ELSE Continue the clustering by doing STEP 4
STEP 7: Final Segmented region was mapped and its area will be calculated.
B. IMPROVED K-MEANS CLUSTERING

Original K-means calculation pick k focuses as starting bunching focuses, various focuses may acquire various arrangements. To reduce the affectability of starting point decisions, we utilize a medoid, which is the most halfway found item in a bunch, to get better introductory focuses. The interest of stochastic inspecting is normally inclination the example to almost address the first dataset, in other words, tests drawn from the dataset can't cause mutilation and can reflect unique information's dispersion. Looking at two arrangements produced by grouping tests drawn from the first dataset and itself utilizing K-implies separately, the area of bunching centroids of these two are practically comparable. Along these lines, the example-based technique is pertinent to refine introductory conditions. To diminish the impact of the test on picking beginning stages, the following methodology is utilized. To start with, drawing various sub-examples (say J) from a unique dataset (the size of each sub-example isn't more than the ability of the memory, and the entirety for the size of J sub-examples is pretty much as close as conceivable to the size of unique dataset). Second, utilize K- implies for each sub-example and delivering a gathering of centroids individually. At long last, contrasting J arrangements and picking one gathering having negligible worth of square- mistake work as the refined introductory focuses.

To try not to separate one major bunch into at least two ones for embracing the square-error rule, we accept the quantity of grouping is K' (K > K, K' relies upon the equilibrium of grouping quality and time). As a rule, greater K' can grow to look through space of arrangement space, and diminish the circumstance that there is no underlying worth close to some extremum. Hence, re- bunching the dataset through K-implies with the picked introductory conditions would create K.centroids, at that point blending K’ groups (which are closest groups) until the number of groups diminished to k.

Algorithm:

STEP 1: Read the input image
STEP 2: Get the number of Clusters (k’ = 2*k)
STEP 3: Calculate the centroid for each cluster.
STEP 4: Calculate distance between each pixel to centroids
STEP 5: Group pixels based on minimum distance
STEP 6: Count non-zero pixels
STEP 7: Update the centroid Values
STEP 8: Check If (old centroid –new centroid) is minimum It End
ELSE Continue the clustering by doing STEP 4
STEP 9: Final Segmented region was mapped and its area will be calculated.
C. FUZZY C-MEANS CLUSTERING

In fuzzy clustering, each point has a level of having a place with groups, as in fuzzy rationale, instead of having a place totally with only one bunch. In this way, focuses on the edge of a bunch, might be in the group less significantly than focuses on the focal point of the bunch. An outline and examination of various fuzzy clustering calculations are accessible. Any point x has a bunch of coefficients giving the level of being in the kth group wk(x). With fuzzy c-means, the centroid of a bunch is the mean, all things considered, weighted by their level of having a place with the group[13].

\[ C_k = \frac{\sum_x w_k(x)x}{\sum_x w_k(x)}. \]

The level of having a place, wk(x), is connected conversely to the separation from x to the group community as determined on the past pass. It additionally relies upon a boundary m that controls how much weight is given to the nearest focus. The fluffy c-means calculation is the same as the k-means calculation:

- Choose various groups.
- Assign arbitrarily to each point coefficient for being in the bunches.
- Repeat until the calculation has merged (that is, the coefficients' change between two emphases is close to, the given affectability limit).
- Compute the centroid for each group, utilizing the equation above.
- For each point, process its coefficients of being in the groups, utilizing the equation above.

The calculation limits intra-group difference too, yet has similar issues as k-means; the base is a nearby least, and the outcomes rely upon the underlying decision of weights. The assumption expansion calculation is an all the more measurably formalized technique that incorporates a portion of these thoughts: incomplete enrollment in classes. Fuzzy c-means have been a vital device for picture preparation in clustering objects in a picture.
Algorithm:

STEP 1: Read the input image

STEP 2: Get the number of Clusters

STEP 3: Initialize the Partition Matrix

STEP 4: Find the Membership matrix

STEP 5: Calculate the centroid Values

STEP 6: Find the Distance between data and centers

STEP 7: Update the Membership values

STEP 8: Calculate the objective function with above values

STEP 9: Check If Objective function difference between two iteration is minimum

    It End

    ELSE Continue STEP 5

STEP 10: Final Segmented region was mapped and its area will be calculated.

VI. SIMULATION RESULTS.

This proposed method is implemented by using MATLAB 2021B
1. **CT LUNG IMAGES:**

![CT Lung Images](image)

**Fig. 5 CT Lung images**

**STEP 1:**

Select an input image taken from CT lung images from 200 CT images database for project purpose initiated by 4 test images.

**STEP 2:** 1. a) Enter the number of corresponding segmentation method.

![Segmentation Selection](image)

**Fig. 6 Input image selection**

After entering the corresponding segmentation as 1, cluster will be formed accordingly.

![Segmentation Type Selection](image)

**Fig. 7 Segmentation Type Selection For KM Clustering**
A) first clustering of K-Means algorithm

(B) Output of second clustering of K-Means algorithm

(C) Output of third clustering of K-Means algorithm

(D) Output of fourth clustering of K-Means Algorithm

2.a) Enter the number of corresponding segmentation method
b) After entering the corresponding segmentation as 3, cluster will be formed accordingly.

(A) Output of first clustering of Improved K-Means algorithm (B) Output of second clustering of Improved K-Means algorithm (C) Output of third clustering of Improved K-Means algorithm (D) Output of fourth clustering of Improved K-Means algorithm

Fig. 11 Segmentation Type Selection For FCM Clustering

3. a) Enter the number of corresponding segmentation method.
3. b) After entering the corresponding segmentation as 2, cluster will be formed accordingly.

(A) Output of first clustering of Fuzzy C-Means algorithm (B) Output of second clustering of Fuzzy C-Means algorithm (C) Output of third clustering of Fuzzy C-Means algorithm (D) Output of fourth clustering of Fuzzy C-Means algorithm

STEP 3: Parameters such as Entropy values are calculated of various methods.

Comparison based Entropy

<table>
<thead>
<tr>
<th>S.No</th>
<th>Test images</th>
<th>Clusters</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>K-Means</td>
</tr>
<tr>
<td>1</td>
<td>Test1</td>
<td>4</td>
<td>0.5494</td>
</tr>
<tr>
<td>2</td>
<td>Test2</td>
<td>4</td>
<td>0.5420</td>
</tr>
<tr>
<td>3</td>
<td>Test3</td>
<td>4</td>
<td>0.4535</td>
</tr>
<tr>
<td>4</td>
<td>Test4</td>
<td>4</td>
<td>0.4421</td>
</tr>
</tbody>
</table>

Fig. 13 Entropy evaluation

Fig. 14 Graph for entropy evaluation
STEP4: FCM based Segmentation Results

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Test Images</th>
<th>Clusters</th>
<th>FCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Test1</td>
<td>4</td>
<td>90.87, 99.82, 99.58</td>
</tr>
<tr>
<td>2</td>
<td>Test2</td>
<td>4</td>
<td>99.70, 99.81, 99.81</td>
</tr>
<tr>
<td>3</td>
<td>Test3</td>
<td>4</td>
<td>97.32, 99.85, 99.87</td>
</tr>
<tr>
<td>4</td>
<td>Test4</td>
<td>4</td>
<td>82.45, 99.83, 99.49</td>
</tr>
</tbody>
</table>

Fig. 15 Results for FCM based Segmentation

Fig. 16 Evaluation measure for FCM based segmentation

Fig. 17 Evaluation measure for FCM based Segmentation

Various Images are taken and segmented using Fuzzy C- Means clustering. Depending on the number of cluster it will perform the operation. The parameters such as Entropy of various methods and sensitivity, specificity and accuracy of FCM are calculated. While comparing the parametric values of the entire algorithm it is shown that Fuzzy C-Means Clustering is the preferable one so evaluated the segmentation.

VII. CONCLUSION

K-Means is useful for several reasons, conceptual simplicity, and ease of implementation. It works best when the dimensionality of the data is not too large and hence requires much iteration. The edges will not detect properly in case of K- Means, whereas it will overcome by updating the centroids with improved K-Means. FCM gives better result for low contrast images. Data point assigned membership to cluster center because of which belong to more than one cluster. The clustering regions will be changed in consecutive execution with FCM.

VIII. FUTURE WORK

K-Means, Improved K-Means, and Fuzzy C-Means are the most famous clustering algorithms. K-Means calculation is utilized for division of CT lung and isn't reasonable for all lighting states of pictures. Interaction time is more with Improved K- Means. FCM simply thinks about the force of the picture and in uproarious pictures, power isn't trustful. Accordingly, this calculation has a bad outcome in low difference, in-homogeneity, and boisterous pictures. Numerous calculations acquainted with make FCM powerful against commotion however by and by the greater part of them were and are perfect somewhat. Once in a while, due to in-homogeneity, low difference, commotion, and disparity of substance with semantic, programmed strategies neglect to fragment the picture effectively. Consequently, for these pictures, it is important to utilize client help to
address the strategy's error. Be that as it may, strong programmed techniques can be created in which client help is limited. At the point when client help is important, the division would be regulated. The diminishing grouping mistake is the benefit of administered strategies and the requirement for client help is the inconvenience of these techniques. Future work is to supplant with the new calculation which distinguishes lung Corona virus more precise than these strategies

REFERENCES