Retraction

Retraction: Efficient Clustering of Unlabeled Brain DICOM Images based on similarity (J. Phys.: Conf. Ser. 1916 012017)

Published 23 February 2022

This article (and all articles in the proceedings volume relating to the same conference) has been retracted by IOP Publishing following an extensive investigation in line with the COPE guidelines. This investigation has uncovered evidence of systematic manipulation of the publication process and considerable citation manipulation.

IOP Publishing respectfully requests that readers consider all work within this volume potentially unreliable, as the volume has not been through a credible peer review process.

IOP Publishing regrets that our usual quality checks did not identify these issues before publication, and have since put additional measures in place to try to prevent these issues from reoccurring. IOP Publishing wishes to credit anonymous whistleblowers and the Problematic Paper Screener [1] for bringing some of the above issues to our attention, prompting us to investigate further.

[1] Cabanac G, Labbé C and Magazinov A 2021 arXiv:2107.06751v1

Retraction published: 23 February 2022
Efficient Clustering of Unlabeled Brain DICOM Images based on similarity

Suriya Murugan¹, M G Sumithra², M Murugappan³
¹CSE, Assistant Professor, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai - 600 062, Tamil Nadu, India.
²Professor, KPR Institute of Engineering and Technology, Coimbatore - 641 407, Tamil Nadu, India.
³Professor, Kuwait College of Science and Technology, Kuwait – 13133. suriya13.ms@gmail.com

Abstract. Clustering has proven to be an effective method in the medical field for finding patterns in labelled and unlabelled datasets. This work is implemented over whole body CT scans (~1TB) of 3500 patients in form of unlabelled DICOM images. The whole-body CT images have been anonymized for 30 attributes based on DICOM regulations and the Brain images alone are segmented using the DICOM tag element called ‘Protocol stack’. The segmented Brain images are efficiently grouped based on visual similarity using K-means clustering after performing feature extraction and dimensionality reduction. The results of the clustering can be further utilized by radiologists to perform labelling or find patterns in Brain CT scans of patients that are difficult where each scan consists of a varying number of slices during detection of Internal Bleeding. The efficiency of K-means is analyzed by performing computation over a different number of clusters (K) by applying silhouette scores to find optimal cluster.

Keywords: Machine Learning, Image Clustering, DICOM, K-means algorithm, Principal Component Analysis (PCA), Silhouette Analysis.

1. Introduction

In a number of areas, technical advances in medicine have had a major effect on the quality of diagnosis and patient care. The medical quality of treatment The Digital Imaging and Communications in Medicine (DICOM) standard is a solution to the problem of medical data interoperability, connectivity, and management [1]. Every medical image in this format has a pixel-value map for the image itself, as well as a variety of standardised tags that are either generated automatically by the acquisition system or manually annotated by physicians [2].

1.1 Overview of Medical Imaging

Medical Imaging involves the use of imaging techniques and procedures to get pictures of the human body that can help in patient diagnosis and care. It can also be used to monitor any ongoing problems, and therefore can assist with recovery plans [3]. Types of medical imaging methods utilize different technologies to capture the images. In recent years, clinical diagnosis based on medical imaging methods gained major attention in patients’ healthcare, both in patient treatment in hospital [4]. Medical imaging consists of a collection of processes or procedures to establish visual images of the body's inner parts...
such as organs or tissues for therapeutic purposes to monitor health conditions, diagnose, and treat diseases and injuries [5].

Particle emission tomography (PET), single-photon emission computed tomography (SPECT), Magnetic resonance imaging (MRI), computed tomography (CT), ultrasonic imaging (US), and planar x-ray devices are some of the other methods used (digital, analogue and portable) and hybrid imaging systems. In this work, the type of modality used was CT scanner and The DICOM files collected from a CT scanner are stored with an extension of ‘.dcm’ and these are visualized and processed by any DICOM visualizer as shown in Figure 1.

![Figure 1. Medical imaging Modality - CT scannerand DICOM image storage system](image)

1.2 Medical Imaging standard - DICOM

DICOM is a common protocol used in many healthcare facilities for storing and transmitting medical images and other clinical diagnosis-related data. DICOM was created by the National Association of Electrical Manufacturers (NEMA) and the American College of Radiology (ACR). It is a registered NEMA trademark and is regulated by the DICOM Standards Committee, a consortium of users across all medical imaging specialties with an interest in standardizing medical imagery information. The DICOM specification encompasses both the formats to be used to store digital forms of medical images and related data, as well as the protocols to be implemented to incorporate various communication systems that are useful in the workflow of medical imaging.

DICOM has become the de facto standard in medical imaging: today the vast majority of all major vendors’ digital medical imaging systems (including acquisition instruments, diagnostic workstations, libraries, servers, medical printers, etc.) adopt and comply with portions of the DICOM standard, depending on the services they introduce. DICOM has also been generally recognized and adopted by medical institutions including public and private hospitals, testing centers, and laboratories of different sizes for research [6].

MicroDICOM viewer is used for displaying the DICOM images in this work. MicroDicom is an application in DICOM format for the primary processing and preservation of medical images. MicroDicom DICOM viewer is built with the most common DICOM image manipulation software and features an intuitive user interface. A snapshot of a part of the brain DCM file shown in Figure 2 is viewed using the visualizer. The viewer is most users friendly and displays the DICOM image along with DICOM tags which includes patient information and other tags [7].

![Figure 2. MicroDicom Viewer](image)
A DICOM tag has the following fields [13]:

- A tag that defines the attribute, usually in hexadecimal format (XXXX, XXXX) and may be further divided into DICOM group number and DICOM item number.
- A DICOM Value Representation (VR) describes the attribute value type and format.

Sample Tag information obtained from the DICOM viewer is given in following Figure 3.

![Figure 3. DICOM Tags](image)

In the DICOM model depicted in Figure 4, A patient should have 1 ..n tests (sometimes called assessments or procedures). Each analysis is composed of series 1 .. n. A series is usually equated with a particular form (modality) of data, or a patient's location on the acquisition unit. Each sequence includes instances of 1 ..n DICOM objects (most commonly images, but also records, artifacts in the waveform, etc.). Each DICOM object in a study contains all this material. Therefore, if a study is conducted on a patient with 2 sets, each with 10 instances, all the instances should contain information about the patient and the study in its header. Also, the instances will contain information about the series it is in, as well as information about its case [8].

![Figure 4. DICOM Image Acquisition Model](image)

1.3 Machine learning for Medical Imaging

The application of Artificial Intelligence (AI) is playing a significant role in healthcare-related equipment design, automated clinical diagnosis, is playing a significant role in healthcare. With imaging
enabled specialties like radiology, the machine learning domain is changing many areas of healthcare delivery. The machine learning domain is transforming many aspects of healthcare delivery, with imaging-enabled specialties such as radiology. In the coming years, medical imaging professionals will have a rapidly expanding AI-enabled diagnostic toolkit at their disposal, to support with all aspects of image interpretation from detection, classification, and segmentation, through to the extraction of quantitative imaging features [9]. Coupled with advancements in Information Technology (IT) infrastructure and the availability of affordable Graphical Processing Units (GPU) based computing and high-performance data storage, the pace of product development for AI-based medical image analysis is faster than ever before. Much of the initial focus for the application of machine learning in medical imaging has been on image analysis and developing tools to make radiologists more efficient and productive.

Similar tools can also help with more effective diagnosis and recovery preparation, as well as reducing missed diagnoses, resulting in better patient outcomes. Beyond clinical decision-making, AI and machine learning have a major influence in the area of radiology. It also assists in enhancing the patient's experience in the imaging process, from the initial image review scheduling to the final diagnosis and follow-up. Machine learning during image reconstruction can improve the quality of low-dose CT scans to that of standard-dose CT scans, reducing patient radiation exposure even further. During image recognition one of the major challenging steps is the process of labeling the data which is the initial process for image construction. All the medical image data are defined using standardized radiological definitions and formats in to implement machine learning algorithms due to data labelling and order compliance within the dataset. Medical image analysis faces many challenges when machine learning is applied and mainly in viewing radiology images [10].

The effectiveness of machine learning in medical image analysis is hampered by the following challenges as depicted in Figure 5

- According to IBM estimations, images currently account for up to 90% of all medical data, i.e., the amount of imaging volumes continue to increase by 5-10% annually. And a major challenge in medical imaging AI deployments is its successful storage. Radiologists and researchers will also aim to make advances in the use of advanced clinical technology, personalized diagnostic and prognostic resources and the ability to improve individual patient results and leverage Big Data in imagery.

- The development of the machine learning model requires all the training data to be in the labelled format, but mostly the medical images are unlabelled due to the in-accessibility of domain expert(s). This process is expensive and inefficient, therefore, often unable to produce enough labels.

- All models of machine-learning are trained on simple 2D images (spatial coordinates). But typically, the CT and MRI images are 3D, basically adding dimension(time) to the issue that needs more experience to manage these types of images [11].

Figure 5. Challenges in Medical Image Analysis
This paper discusses the possibility of automatically clustering unlabelled brain DICOM images from a massive big data of nearly 1 TB dataset and clustering the brain images based on similarities between image features. Using the proposed algorithm, we try to demonstrate that clustering of DICOM image results in generating visually identical clusters. The outcome of this work can be used by radiologists for further labelling in a fast and efficient manner and further which can be utilized in the development of a deep learning architecture for automatic feature identification of Intracranial haemorrhages (ICH).

The work is structured as follows. Section 2 gives an overview of image clustering and its types. The source of the dataset and its specifications are described in Section 3. The proposed system architecture of this present work is explained in section 4 followed by performance analysis along with experimental results in section 5 and the conclusion is given in section 6.

2. Overview of Image Clustering

Clustering is a crucial data mining activity that aims to classify the data objects of different clusters based on their similarities. Image clustering is used for image extraction and pixel segmentation with data objects being images. Clustering preserves two properties namely homogeneity and heterogeneity optimization between intra and inter clusters respectively.

Clustering is an unsupervised learning method, which is capable of handling unlabelled data and image clustering involves mapping image groups within each cluster based on similarity. It is much used in image comprehension of features in image like colour, texture and shape or can be semantic. Similarity computation functions must be powerful to get accurate results. This also helps to extract the smaller volume of data from a large dataset. It is used primarily for image segmentation and Content-Based Image Retrieval (CBIR).

The clustering techniques for any image mainly follow two stages, as shown in Figure 6. First, estimation of the measure of similarity using feature extraction. The second step is the implementation of a clustering algorithm. Clustering employs a similarity matrix to establish the degree of similarity between two image segments. Since the image data is increasing, when the given image is used for similarity computation, this computational operation becomes repetitive. Feature extraction significantly decreases the cost of similarity computation. During this process the device accepts image data as input and generates image cluster as output [12].

![Figure 6. Overview of Image Clustering](image)

2.1 Types of Medical Image clustering

The field of medical imaging plays an important role in healthcare where clustering helps to perform image segmentation in numerous ways, e.g., recognition of diseases, organ-detection etc. The clustering method is highly useful in managing massive medical huge databases of medical images is highly complex and it contains a variety of features [13].
Figure 7. Types of Image Clustering

Figure 7 depicts the types of various unsupervised clustering algorithms [14] that can be used for image clustering and each one of them is explained below:

- **Hierarchical clustering** is a process by where image clusters are built by recursive portioning instances. Techniques from the top-down or bottom-up methods are used to perform hierarchical clustering. All instances begin in one cluster in divisive hierarchical clustering, and this cluster is iteratively divided into sub clusters until the desired number of clusters is reached.

- **The Partitioning clustering** technique moves cluster instance from one cluster to another hence it is necessary to predefine the number of clusters. Squared error clustering methods are the most commonly known approaches of partition clustering. One such approach is the K-means clustering method, which is applied in most clustering applications of medical images where the dataset is large and unlabelled.

- **Density Methods** are primarily used to discover clusters of arbitrary form by considering that the points belonging to a cluster are derived from common distribution. One method based on density is the DBSCAN method which searches the neighbourhood object's in the database and checks whether it has least amount of discoverable cluster objects.

- **Model-based approach** aims at optimizing the fitness among data and mathematical processing. The data is represented as a hierarchical structure that has prototypes for each cluster and neurons representing the input and where the neurons are bound to the prototype neurons.

- **Grid-based approach** is primarily used to performs faster processing by splitting the space into number of cells and then conducting clustering activity.

- **Fuzzy clustering** in this all the data points are allocated to a cluster space using membership value such that one or more clusters can be allocated to each data point.

3. Data Collection

The dataset for this project is CT scan images of patients of nearly 3500 patients and these includes 11,00,000 whole body CT scan images and each scan consist of varying number of series. All input images are in dicom from the whole-body CT scan dataset, with proper anonymization, the brain part is alone segmented for further processing which consists of 1,25,000 images and is outlined in the following Table 1.

| Table 1. Dataset Specification |
|--------------------------------|
| Total no of patients            | 3500                          |
| Type of Modality                | CT scan                       |
| Image format                    | DICOM(.dcm)                   |
| Labelling                       | Unlabelled                    |
| Number of Whole-Body CT Images  | 11,00,000                     |
Total Brain CT scan images | 1,25,000

4. Proposed Architecture Model

The objective of the work is to efficiently and automatically segregate brain images from whole-body image data archive. Since these are unlabelled dicom images and since the cost of labelling is so high which requires trained radiologists, the automatic grouping by making the system learn from feature vectors by considering all pixels of the image to perform clustering process. The steps involved in the work starting from collection of dataset to grouping of brain images based of similarity is outlined in Figure 8 and each step is explained below.

**Figure 8. System Architecture**

4.1 Anonymization

As mentioned in Section 1.2 and Figure 2, patient demographic information and a host of other image research information are encoded under the DICOM tags within an image header. It is necessary to protect the privacy of the patient while the medical images are used in data analysis. This is implemented by following the “Anonymization” process, whereby confidential information about the patient is removed or renamed from the DICOM header when a DICOM file is uploaded for processing. Among the entire 97 DICOM tags, the whole-body CT images (~1TB) of 3500 patients have been anonymized for 30 attributes based on DICOM regulation standards. For example, ‘PatientID’ is anonymized as ‘AHXX’, ‘BHXX’ and so on (the first 2 strings are the alphabet and the next 2 strings with random number).

4.2 Segmentation

Image segmentation is a mechanism where an image is partitioned into many sub-regions. Segmentation is a crucial phase in the processing of images, and medical images have properties that can render them difficult to segment, such as various tissues with identical intensities, organs that occlude and overlap with each other, and individuals with significant physical variation. In this work the Brain images alone are segmented from the whole-body CT scans using the attribute called ‘Protocol stack’ from DICOM tags and collected to a separate file.

4.3 Feature Extraction and Selection

Extracting meaningful information about the characteristics of medical images is highly essential for developing an intelligent autonomous diagnosis system. This information is referred to as ‘features’ and it is useful for pattern recognition. This stage is about the processing of image character quantization.
After extracting the features, the selection of subset of the most robust element is performed to identify the similarity between images and thus the overall complexity [15].

![Image](image1.png)

**Figure 9. Feature Extraction and Selection**

As depicted in Figure 9, a two-step process is performed to extract salient features:

- **Feature Extraction**
  
  In this step, all the DICOM brain images of size 512x 512 are fed as input to the feature extraction model to generate 2, 62,144 feature vectors as output.

- **Feature Selection**
  
  To reduce the size of features i.e, to overcome the curse of dimensionality, Principal Component Analysis (PCA) is utilized which aims to retain the essential components that have more data variance and eliminate the non-essential components with less variability. It can cluster related data points without any marks, based on the similarity of features between them. PCA transforms a matrix of ‘n’ functions into a new dataset with features lower than n. From 2,62,144 feature vectors, the algorithm has identified and generated reduced feature vectors based on image similarity and is shown in Figure 10.

![Image](image2.png)

**Figure 10. Reduced feature space using PCA**

### 4.4 Clustering using K-Means algorithm

The k-means is an unsupervised learning technique and this is the most preferred clustering method for clustering a huge size of unlabelled data. The main goal of this algorithm is to divide the data points in a data set into different categories or groups. The data points are grouped together based on their similarities. K-means tries to partition the data set into k-clusters using an objective function based on the principal components obtained from PCA.

When k spatially dissimilar cluster group are chosen, data is separated using Euclidian Distance for producing successful mining output. Each cluster group contains a core called the centroid and depending on how similar the features are to the centroid, a data point is clustered into a cluster. Let \( X = \{x_1, x_2, x_3, \ldots, x_n\} \) represent data points and \( V = \{v_1, v_2, \ldots, v_c\} \) represent cluster centres.
Then new cluster center is calculated as in Eqn 1:

\[ v_i = \left( \frac{1}{c_i} \right) \sum_{j=1}^{c_i} x_i \]  

(1)

Here, 'c_i' denotes the total number of data points in the i\(^{th}\) cluster.

The process is repeated with a different number of ‘k’ values (k=2 till k=15). Figure 11 (a) depicts the sample of clusters when ‘k=8’ and Figure 11 (b) is the result of similar images grouped under cluster five of k=8.

5. Performance analysis and result

In this work, the Silhouette score has been used for assessing the performance of k-means clustering. This score has been most widely used in the literature for clustered datavisualization and assessment [15]. The silhouette score contains details about the degree to which a particular object fits into a specific cluster, taking into account cluster tightness and separation. The Silhouette score ranges from +1 to -1, with a high value approaching +1 indicating that the data point is important and grouped in correct cluster.

The Silhouette score is computed as s(i) for each data point i and is given in Eqn 2:

\[ s(i) = \begin{cases} \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, & \text{if } |C_i| > 1 \\ 0, & \text{if } |C_i| = 1 \end{cases} \]  

(2)

Here, \( a(i) \) denotes the similarity of i to its cluster and is measured as the average distance of i from other points in the cluster and is calculated as in Equ 3.

For each i in C_i:

\[ a(i) = \frac{1}{|C_i|-1} \sum_{j \neq i \in C_i} d(i,j) \]  

(3)

Similarly, \( b(i) \) given in Equ 4 indicates the measure of dissimilarity of i from points in other clusters.

\[ b(i) = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \neq i} d(i,j) \]  

(4)
Here $d(i, j)$ is the distance between points $i$ and $j$ which is calculated based on the Euclidean Distance as in Table 2.

**Table 2. Silhouette Score for different ‘k’**

| No of Clusters (k) | Silhouette Score (s(i)) |
|--------------------|-------------------------|
| 2                  | 0.453313172             |
| 3                  | 0.473416371             |
| 4                  | 0.513564332             |
| 5                  | 0.518649398             |
| 6                  | 0.573226036             |
| 7                  | 0.588591393             |
| 8                  | 0.595128428             |
| 9                  | 0.589333717             |
| 10                 | 0.588333717             |
| 11                 | 0.600986188             |
| 12                 | 0.586416810             |
| 13                 | 0.575966755             |
| 14                 | 0.575097614             |
| 15                 | 0.576033868             |

Generally when the Silhouette score values are greater than 0.50, then it results in better cluster structures. In this work, since the evaluation is based on visual similarity only for brain images the cluster range was fixed from 2(based on a number of series) to a maximum of 15. The plotting of different k values and their silhouette score is shown in Figure 12.

**Figure 12. Silhouette Analysis to find optimal ‘k’**

For $k=2$, the score was $s(i) = 0.453313172$, and when the cluster value incremented the corresponding silhouette score also increased nearing to +1. But after a threshold point say $k=12$, the score started decreasing gradually. This makes clear that based on the feature vector the maximum number of similarity grouping can be minimum $k=8$ or maximum $k=11$ clusters. Compared to both $s(i) = 0.600986188$ is maximum, hence $k=11$ can be selected as optimal cluster value for pixel-based similarity clustering of brain images.
6. Conclusion
Medical image analysis is a challenging task when the (a) dataset is huge and (b) particularly when all the images are unlabelled. This requires assistance of trained radiologists for further processing whether it is to identify the type of organ, presence of disease or location of disease to take necessary medical action. The data collected for this project was to perform an automatic Haemorrhage detection system. The radiologists have to examine a large stack of whole-body scans and thus the work load may lead to potential error or delay which may adversely affect disease diagnosis for patients. In order to overcome the challenges stated, the whole body DICOM images are segmented to extract only the Brain part using DICOM tag attribute. Then feature extraction and dimensionality reduction is applied to create a feature vector to retain all necessary pixels and reduce the curse of dimensionality (2,62,144) using the PCA technique. The images are grouped efficiently using the K-means clustering algorithm and using silhouette score analysis the efficiency of the cluster value is found to be k=11. The result of this process can assist the radiologists for further labelling by selecting the necessary brain slices (folders) and also helpful for further development of a deep learning model to detect and diagnose the extent of brain trauma including brain haemorrhage.

References
[1] Zhang, W.-L., & Wang, X.-Z. (2007). Feature Extraction and Classification for Human Brain CT Images. 2007 International Conference on Machine Learning and Cybernetics. doi:10.1109/icmlc.2007.4370318.
[2] Chen Jia-lin, He Hua-can, & Liu Cheng-xia. (2009). Feature extraction of brain CT image based on target shape. 2009 Chinese Control and Decision Conference. doi:10.1109/ccdc.2009.5192600.
[3] Wazarkar, S., & Keshavamurthy, B. N. (2018). A survey on image data analysis through clustering techniques for real world applications. Journal of Visual Communication and Image Representation, 55, 596–620. doi:10.1016/j.jvcir.2018.07.009.
[4] Arbabshirani, M. R., Fornwalt, B. K., Mongelluzzo, G. J., Suever, J. D., Geise, B. D., Patel, A. A., & Moore, G. J. (2018). Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. Npj Digital Medicine, 1(1). doi:10.1038/s41746-017-0015-z.
[5] Anh-CangPhan, Van-QuyenVo&Thuong-CangPhan (2019) A Hounsfield value-based approach for automatic recognition of brain haemorrhage, Journal of Information and Telecommunication, 3:2, 196-209, DOI: 10.1080/24751839.2018.1547951.
[6] Haldorai, A. Ramu, and S. Murugan, Social Aware Cognitive Radio Networks, Social Network Analytics for Contemporary Business Organizations, pp. 188–202. doi:10.4018/978-1-5225-5097-6.ch010
[7] R. Arulmurugan and H. Anandakumar, Region-based seed point cell segmentation and detection for biomedical image analysis, International Journal of Biomedical Engineering and Technology, vol. 27, no. 4, p. 273, 2018.
[8] Dawud, A. M., Yurtkan, K., & Oztoprak, H. (2019). Application of Deep Learning in Neuroradiology: Brain Haemorrhage Classification Using Transfer Learning. Computational Intelligence and Neuroscience, 2019, 1–12. doi:10.1155/2019/4629859.
[9] Aiello, M., Cavaliere, C., D’Albore, A., & Salvatore, M. (2019). The Challenges of Diagnostic Imaging in the Era of Big Data. Journal of Clinical Medicine, 8(3), 316. doi:10.3390/jcm8030316.
[10] Lee, C. H., & Yoon, H.-J. (2017). Medical big data: promise and challenges. Kidney Research and Clinical Practice, 36(1), 3–11. doi:10.23876/j.krcp.2017.36.1.3.
[11] Manojlović, T., Ilić, D., Miletić, D., & Štajduhar, I. (2020). Using DICOM Tags for Clustering Medical Radiology Images into Visually Similar Groups. Proceedings of the 9th
International Conference on Pattern Recognition Applications and Methods. doi:10.5220/0008973405100517.

[12] Zhang, Y. C. and Kagen, A. C. (2017). Machine Learning Interface for Medical Image Analysis. Journal of Digital Imaging, 30(5):615–621.

[13] Kaufman, L. and Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics).

[14] Gueld, M. O., Kohnen, M., Keysers, D., Schubert, H., Wein, B. B., Bredno, J., and Lehmann, T. M. (2003). Quality of DICOM header information for image categorization. In Medical Imaging 2002: PACS and Integrated Medical Information Systems: Design and Evaluation, 4685, pages 280–287. SPIE.

[15] Rahman, M. M., Bhattacharya, P., and Desai, B. C. (2007). A Framework for Medical Image Retrieval Using Machine Learning and Statistical Similarity Matching Techniques With Relevance Feedback. IEEE Transactions on Information Technology in Biomedicine, 11(1):58–69.

[16] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics.