Using DICOM Tags for Clustering Medical Radiology Images into Visually Similar Groups

Manojlović, Teo; Ilić, Dino; Miletić, Damir; Štajduhar, Ivan

Download date / Datum preuzimanja: **2025-03-26**

Using DICOM Tags for Clustering Medical Radiology Images into Visually Similar Groups

Teo Manojlović[1] [a], Dino Ilić[1], Damir Miletić[2] [b] and Ivan Štajduhar[1] [c]

[1]*University of Rijeka, Faculty of Engineering, Department of Computer Engineering, Vukovarska 58, 51000 Rijeka, Croatia*

[2]*University of Rijeka, Clinical Hospital Centre Rijeka, Clinical Department for Radiology, Krešimirova 42, 51000, Rijeka, Croatia*

Keywords: PACS, DICOM, Medical Imaging, Visual Similarity, Clustering, K-medoids.

Abstract: The data stored in a Picture Archiving and Communication System (PACS) of a clinical centre normally consists of medical images recorded from patients using select imaging techniques, and stored metadata information concerning the details on the conducted diagnostic procedures - the latter being commonly stored using the Digital Imaging and Communications in Medicine (DICOM) standard. In this work, we explore the possibility of utilising DICOM tags for automatic annotation of PACS databases, using K-medoids clustering. We gather and analyse DICOM data of medical radiology images available as a part of the RadiologyNet database, which was built in 2017, and originates from the Clinical Hospital Centre Rijeka, Croatia. Following data preprocessing, we used K-medoids clustering for multiple values of K, and we chose the most appropriate number of clusters based on the *silhouette score*. Next, for evaluating the clustering performance with regard to the visual similarity of images, we trained an autoencoder from a non-overlapping set of images. That way, we estimated the visual similarity of pixel data clustered by DICOM tags. Paired t-test ($p < 0.001$) suggests a significant difference between the mean distance from cluster centres of images clustered by DICOM tags, and randomly-permuted cluster labels.

1 INTRODUCTION

The advances of technology in medicine directly influenced the quality of diagnosis and treatment of working with patients in numerous fields. Most notably, these involve new methodologies and techniques, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). In addition to new medical procedures and techniques, advances in technology allowed medical personnel to create, store and retrieve various information about patients, medical images and other relevant data. To accommodate for all of that, Picture Archiving and Communication Systems (PACS) have been developed with a goal of providing efficient storage, fast retrieval of data and many other features (Choplin et al., 1992). Even with this system in place, there was a need for a systematic and organised way of transferring medical images between different devices. Digital Imaging and Communications in Medicine (DICOM) [1] is a standard that solves the problem of interoperability, commu-

nication and managing of medical data. In this standard, each medical image consists of a pixel-values map for the image itself, and of a number of different structured tags, either automatically generated by the acquisition device, or manually set by physicians (Bidgood et al., 1997). The amount of data stored by an average clinical radiology department is rapidly increasing each year. Subsequently, finding similar cases and navigating PACS repositories, containing vast amounts of data, becomes more difficult. A suitable annotation of medical images would make this process easier, however manual annotation of images is expensive and time-consuming (Dimitrovski et al., 2011). On the other hand, automatic clustering of data contained in a PACS would allow for easier navigation and exploration of similar cases. Hence, this paper explores the possibility of automatic clustering of medical images based on the assigned DICOM tags. This study involves the analysis of the contents of a sizeable PACS dataset, and the exploration of the performance of an unsupervised machine learning technique, i.e. K-medoids, for automatic clustering of data.

Because the data stored in the relational database of Clinical Hospital Centre (CHC) Rijeka is insuffi-

----

[a] https://orcid.org/0000-0002-8891-0935
[b] https://orcid.org/0000-0003-2890-1890
[c] https://orcid.org/0000-0003-4758-7972
[1]https://www.dicomstandard.org/

ciently informative for the clustering of images, such that they would have a sufficient level of detail image-content wise, DICOM tags associated with the images were used for this purpose. DICOM files contain numerous meta tags that were recorded during exams (approximately 4,000 tags total, over the entire dataset). These tags are structured according to their function, which does not necessarily hold for their contents. The tags were first analysed both functionally (manually) and content-wise (manually and automatically), in order to determine the most effective ways of extracting features, to provide good clustering quality.

A specific goal of our clustering-based annotation is to create a quality foundation for training a deep convolutional neural network (CNN) (Krizhevsky et al., 2012) for transfer learning in medical radiology imaging (Qiang Yang and Pan, 2010; Yosinski et al., 2014). When ground-truth labelling exists, e.g. (Xie et al., 2016; Yang et al., 2010), clustering performance is normally evaluated in a supervised manner (e.g. classification accuracy, F1 score, and so on). In our case, because ground truth was unavailable, an alternative (unsupervised) evaluation approach was used instead, using an image feature-extraction mechanism, i.e. an embedding. Because cross-domain embedding (e.g. adapted ImageNet pretrained model) was shown to be inadequate for evaluating our clustering examples, we opted for building an embedder from the data sampled from the same distribution. Therefore, we propose a clustering-evaluation pipeline for handling cross-domain unlabelled data. We elaborate on how this pipeline can be used for an unbiased estimate of visual similarity between images. By using the proposed evaluation, we show that DICOM clustering results in creating visually similar clusters. We thus demonstrate that DICOM clustering produces embedded image representations that are statistically better than random groupings. This proves that the information contained within the DICOM labels is useful for grouping visually similar images.

This work is structured as follows. First, we describe the related work concerning the potential and applicability of DICOM tags for various tasks, also categorising database data. Next, in section 2, we present the dataset used for clustering using DICOM tags, as well as the data-analysis and clustering tools utilised for organising the data. In section 3, we describe the experimental setup, involving also the description of the clustering-evaluation pipeline, along with the results of the evaluation. Finally, in section 4, we present a summary of the results obtained, and give a conclusion.

1.1 Related Work

Although the research focusing on DICOM tags is sparse, several papers have been published describing the use of DICOM tags and categorisation of medical images. As an example, the quality of specific DICOM headers for image categorisation is presented in (Gueld et al., 2003). Although this research has shown that automated categorisation is implausible, it should be noted that the utilised sample size was relatively small, and only one DICOM tag was used for evaluation. Because the DICOM standard is comprised of both automatically generated, and manually filled in data, the question of using multiple different tags for categorisation arises. In (Rahman et al., 2007), it is argued that due to the size of DICOM images, they are not suitable for a web-based environment, however, they contain very important information which can be used for image retrieval from databases.

Usage of DICOM tags is not limited to categorisation only. Researchers have shown that some of the DICOM tags can be used for optimising dose levels present in detectors (Källman et al., 2009). This example shows how using DICOM tags can improve the workflow in medical procedures while maintaining vendor interoperability. By having access to numerical data, such as presented in (Källman et al., 2009), it becomes obvious that a large number of optimisation methods and machine learning algorithms can be applied to these data. In (Avishkar Misra et al., 2005), a C4.5 model was trained with the help of lung shape features and DICOM tags to predict lung regions, attaining highest accuracy for apical region (96.6%) and lowest for middle region (92.5%). DICOM images that are properly stored and managed can provide valuable information for later studies. A well developed image-retrieval system will enable researchers and medical personnel to query patients having similar medical conditions. Proof of concept applications were developed in (Van Soest et al., 2014) to store DICOM metadata in an RDF repository, and to calculate imaging biomarkers based on DICOM images, which enabled searching for images having similar tumour volumes. In (Zhang and Kagen, 2017), an artificial neural network (ANN) was trained by extending the TensorFlow API to process raw DICOM images of basal ganglia, achieving 93.8% accuracy on the classification task where the main goal was to detect if a patient suffers from Parkinson's disease.

Disadvantages of the DICOM standard include rough structuring, ambiguity, and often-optional fields. All of that makes the problem of categorising medical images difficult. To mitigate the afore-

mentioned problems, a new classification code called IRMA was proposed in (Lehmann et al., 2003), having some advantages with regards to content retrieval systems used in medical applications, compared to the DICOM standard. Another approach to medical image annotation was presented in (Dimitrovski et al., 2011), using hierarchical multi-level classification. Reasons behind this approach were similar to the ones presented in (Lehmann et al., 2003), and are related to the DICOM-standard drawbacks. Even though in this work a new way of image annotation is provided, it is also stated that automated categorisation of medical images using DICOM tags is highly desirable (Dimitrovski et al., 2011).

2 MATERIALS AND METHODS

To be able to categorise images, the data had to be sourced, analysed and processed, so the models could be properly trained. This section describes the origin of the data, the process used to get the final dataset, and the methods for finding a model that will support semantic (image-content related) clustering of medical images based on DICOM tags. Described methods and tools were implemented in *Python* using *Pandas* for data manipulation, *Matplotlib* for generating visualisations, *scikit-learn pyclustering* for training the models, and most importantly, *Dask* which provided an API over *Pandas*, *Numpy* and *scikit-learn*, which in turn enabled parallelization while retaining most of the functionality from mentioned packages. For training the convolutional autoencoder, *TensorFlow 2* was used.

2.1 RadiologyNet Dataset

Upon receiving a clearance from the legally competent Ethics Committee, an anonymised, sizeable collection of radiology scans was acquired from the CHC Rijeka PACS in 2017, through the project *UNIRI 16.09.2.2.05*. The collection containing approximately 20 TB of data (approximately 30 million images, 2.4 million sequences; 1.3 million exams) was retrieved from the PACS and stored on a workstation in the possession of the Faculty of Engineering in Rijeka (RITEH), for further work. These images were recorded during the past decade in several localities in Rijeka, on multiple devices using several imaging modalities (mostly MRI and CT). The dataset also involves repeat exams. From the original dataset, approximately 14 million images are described by at least one DICOM tag - henceforth, only this subset was considered for performing the study.

The computational power required for analysing these data was overwhelming, nonetheless. Therefore, in order to make the experiments feasible, in this work we consider only a smaller subset of data, consisting of approximately 5% of the dataset, which resulted in a collection of 668849 images randomly sampled from the original dataset. Images (2D slices) belonging to volumes were treated as independent images in this collection. Number of slices per volume varied in size, depending on the exam type and the imaging modality used. Because this sample is still moderately sized, we believe that the conclusions presented here can be considered relevant, with regard to the task at hand.

Each DICOM file contains a 2D image, as well as a set of corresponding tags. Images of slices shaping specific volumes share the same DICOM tags, with the exception of those tags related to the relative slice location. These tags provide useful information about the procedure undertaken, set both automatically (by the machine), and manually (by the operator). The designations of all of these tags are specified by the DICOM standard, as well as value representation and the possible lengths of the field.

2.2 Data Analysis and Processing

To be able to understand the data and what is contained within the dataset, we performed a frequency analysis of tag values. This process consisted of extracting all distinct values for each DICOM tag individually, and calculating the frequency of those distinct values within a specific DICOM tag. Not only did this allow a better understanding of the data, but it was also necessary for determining the number of missing values, and their percentage for each tag.

For a better understanding of the dataset specifics, a couple of DICOM tags are explained in more detail. One of those tags is "BodyPartExamined", which, as the name suggest, contains the information about the anatomic region examined in a specific diagnostic procedure. This tag is particularly interesting because it is manually entered by performing physicians, which introduces noise because of mislabelling, as shown in (Gueld et al., 2003). However, mislabelling of this specific tag, or any other, was not evaluated because it would require additional assistance from experts, and a substantial amount of time because of the size of the dataset used. One of the examples that can be found in the dataset is the similarity between values. E.g., *thoracic spine*, which can be found in "BodyPartExamined" field is written both as "T_SPINE" and "TSPINE", which of course is considered to be the same anatomic region, and had to

be further processed. That goes for multiple different values in this DICOM tag.

Unlike "BodyPartExamined" which is manually entered during an exam, an example of a machine-generated tag is "Modality". This tag stores the information concerning the medical procedure conducted during a specific exam, such as CT or MRI. Being an automatically generated tag means that missing values are non-existent, and the data should not contain any noise from mislabelled data which occurs for manual input.

Regardless of the features being automatically or manually entered into a DICOM file, most of them require some sort of preprocessing in order to make them suitable for information extraction. An example of such tag is "PixelSpacing", which contains the information on physical distance between centres of pixels in a 2D grid. This value is represented as a two-value array encoded as a string in the following format: "$[x, y]$". This is just one example where the data type cannot be inferred because the value is encoded as a string. For this one, as well as many other tags, we had to write preprocessing parsers. This was shown to be a demanding process, and some of the parsers have proven to be somewhat difficult to write. This was one of the main reasons that led to the reduction of the original feature space used in our experiments. The reduced dataset contained 580 features, some of which were easily converted into corresponding data types. However, many of them were impossible to automatically convert, and a notable share contained a large amount of missing values, subsequently making them less informative, and, henceforth, rendering them unsuitable for our goal.

To determine which features to include in the final dataset, shares of missing values were calculated during frequency analysis. These values were sorted in descending order, and only the first 70 features (those containing the smallest shares of missing values) were included. It was possible to include more features, however, this would have required writing additional parsers which would have greatly increased the time needed for data preprocessing. Furthermore, a number of possible features were dropped from the beginning because some of them (e.g. "SOPInstanceUID") were uninformative due to their uniqueness for each record, whereas other (e.g. "ProtocolName"), although rather important, were extremely difficult to parse due to manual input from the operator.

After feature (DICOM tag) selection was performed, some of the selected features required additional parsing and mapping before the dataset could be finalised for clustering. Parsing was performed for features containing arrays of numbers encoded

as strings. This was done because encoded data is meaningless unless transformed to numerical values, which is why these features were expanded into several columns, depending on the number of items in the encoded array. This also means that the reduced dataset (after feature selection), which contained 70 features, was then expanded to a total of 85 features. Regardless of the problem downscaling, which was necessary because of the overwhelming computational load, the used data remained sufficiently informative for providing meaningful clustering of the data in the DICOM space, which we prove experimentally both there and in the autoencoder image-embedded space.

2.3 Clustering

Because the Euclidean distance is not applicable directly for mixed data types, the similarity measure proposed by Gower (Gower, 1971) was used instead. The similarity metric is calculated using the following expression:

$$S_{ij} = \sum_{k=1}^{p} s_k(x_{ik}, x_{jk})\delta_k(x_{ik}, x_{jk}) / \sum_{k=1}^{p} \delta_k(x_{ik}, x_{jk}), \quad (1)$$

where p is the total number of features, s_k is the similarity score between k-th feature of the i-th and j-th data instance. Because there exists a possibility that some variable is not present in the sample, δ factor is calculated in the following way - it equals 0 if the factors are not comparable, and is 1 otherwise. This solves the problem of existing missing values in the data.

For the categorical features, the similarity score between the k-th categorical variables of the i-th and j-th data instance is calculated using:

$$s_k(x_{ik}, x_{jk}) = \begin{cases} 1 & x_{ik} = x_{jk}, \\ 0 & \text{otherwise}, \end{cases} \quad (2)$$

and, for the numerical features, similarity is calculated using:

$$s_k(x_{ik}, x_{jk}) = \frac{|x_{ik} - x_{jk}|}{R_k}, \quad (3)$$

where R_k denotes the range for k-th variable. Finally, the Gower distance between two data instances, i-th and j-th, is calculated using the following expression:

$$\sqrt{1 - S_{ij}}. \quad (4)$$

Although many clustering techniques could be applied to this problem, in this work we chose to use K-medoids (L. and P., 1987). This algorithm belongs to the family of partitional clustering algorithms, and is

similar to *K*-means, albeit having two important differences. Firstly, in *K*-medoids, the initial cluster centres are selected among dataset points, and the absolute distance between cluster centres is minimised. It is important to note that, in terms of accuracy, the K-medoids algorithm is less sensitive to outliers (Park and Jun, 2009). The first step of the algorithm execution starts with *K* greedily selected medoids, and in each iteration each training example is assigned to its closest centroid. After that, for each medoid *m*, the switch between the medoid *m* and the non-medoid *o* is made and cost change is calculated. If the best swap of *m* and *o* decreases the cost function, *m* and *o* are swapped. This process is continued in iterations until convergence is reached.

2.4 Evaluation Metrics

When evaluating clustering results, two main questions emerge. The first question is related to the optimal number of clusters used. A method for visualisation and assessment of cluster numbers, commonly used and very popular, is the *silhouette score* (Rousseeuw, 1987). *Silhouette score* carries the information concerning the extent to which a specific object fits into some specific cluster, taking into consideration the tightness and the separation of clusters. *Silhouette score* values are always falling into the interval $[-1, 1]$, where, the more positive a result is, the better the separation of clusters is. (Kaufman and Rousseeuw, 1990) provides one possible subjective interpretation of the *silhouette score*. For *silhouette score* value smaller than 0.25, we can conclude that there is no substantial structure. Interval $[0.25, 0.50]$ shows that the structure is weak and requires additional analysis. If the *silhouette score* value falls inside the interval $[0.51, 0.70]$, we can conclude that a reasonable structure has been found, and interval $[0.71, 1]$ shows that a strong structure has been found.

The second step in the evaluation process is to examine the visual similarity of the objects clustered together. There are two popular approaches to extract visual features from images in an unsupervised manner. The first approach is to use a pretrained model, e.g. a VGG16 (Simonyan and Zisserman, 2014) model architecture trained on ImageNet dataset. This, however, exhibited low performance, even with additional fine-tuning of model parameters. Instead, an autoencoder was used ((Masci et al., 2011), (Chen et al., 2017)). Autoencoder is an unsupervised neural network used for learning the data encoding, which is most often used for reducing the dimensionality of input data. In our case, it was used to learn a visual embedding. The autoencoder was trained using

a sample of 30000 images which do not appear in the dataset that was used for clustering (a separate non-overlapping subset of the entire dataset was used instead).

All images were resized into resolution 128*x*128 and their pixels were normalised to fall into the range $[0, 1]$. Our trained autoencoder is a convolutional autoencoder consisting of two main parts. The first part is the encoder, consisting of a sequence of two convolutional layers having 64 filters (dimensions 3*x*3 and 2*x*2, respectively), and a max pooling layer (2*x*2 filter having stride 2). The decoder part of the architecture consists of a sequence of two convolutions layers (having 64 filters with kernel size of 2*x*2 and 3*x*3, respectively), followed by an upsampling layer. The last encoder and the first decoder layers have 20 filters. The number of filters is decreased in order to lower the dimensionality of the image embedding. The architecture of the autoencoder used is shown in Figure 1. To train the model, we use the *RMSprop* optimiser, having MSE as the loss function. Training was done for 400 epochs using batch size of 40 images.

After the autoencoder was trained, the encoder part was used to extract relevant image features, to be used for calculating the distance between images. Because the space of the visual features on the one hand, and the DICOM tags, on the other, is inherently different, one cannot expect that the *silhouette score* will match in both domains. Furthermore, if the clusters contain visually similar images, it is less important how close they are to other clusters compared to the scenario where they contain visually diverse images within a specific cluster. For these reasons, we calculated the mean distance from the samples with visual features and cluster centres.

A diagram depicting the entire evaluation procedure is shown in Figure 2.

3 RESULTS

The first step in the evaluation was related to selecting the optimal number of clusters. We sampled a subset of 5000 data instances from the original dataset to perform the clustering. To speed up the computation, we calculated a distance matrix between data instances prior to algorithm execution. The above mentioned subset size is chosen for two reasons. The first reason is the computation time required for the computation of the distance matrix which grows quadratically with the dataset size. Also, since the distance matrix is kept in RAM during the computation, there exists a specific limit in size which can be used. Clustering was done for $K = \{5, 10, 15, 20, 100, 200, 300\}$
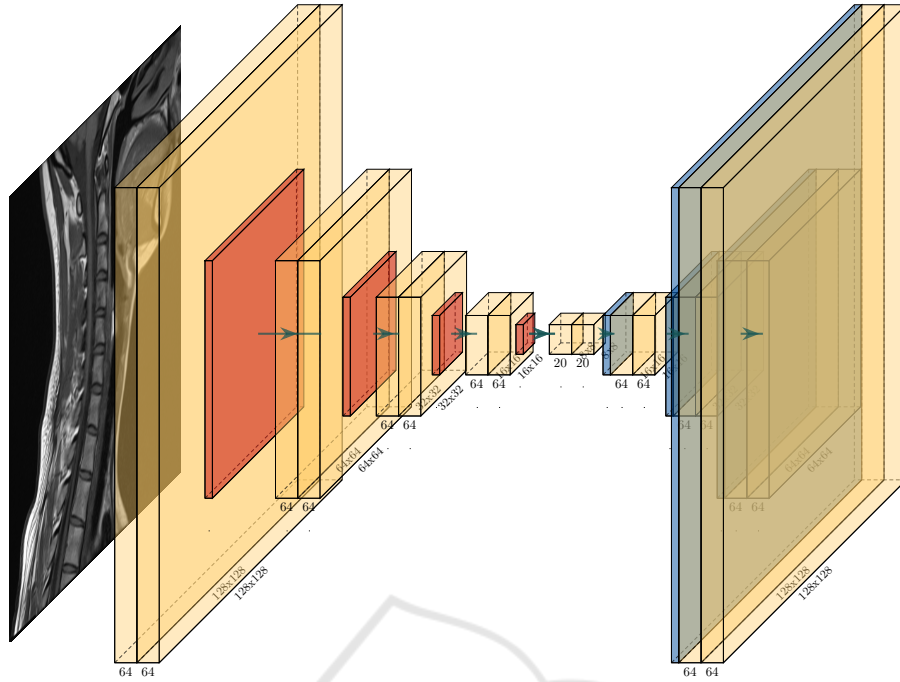
Figure 1: Convolutional autoencoder architecture used for the evaluation of clustering in the pixel space.

clusters, resulting in *silhouette scores*= $\{0.29, 0.32, 0.37, 0.34, 0.31, 0.25, 0.21\}$, respectively. Because the *silhouette score* of 0.37 is the largest, we considered only the cluster size of $K = 15$ for further analysis. *Silhouette score*-per-samples, depicted in Figure 3, shows the existence of several groups which are well clustered. As it can be seen in Figure 3, the samples are sorted within clusters by their *silhouette score*.

The next step in the analysis was to test if the visually similar images are grouped together using the clustering rules inferred from the DICOM tags. After calculating the best number of clusters, we sampled 10 non-overlapping datasets consisting of 5000 DICOM images each, which were then clustered by K-medoids algorithm using $K = 15$ cluster centres. Also, we extracted the visual features from these images using the encoder part of the previously described autoencoder. By using the visual-feature embeddings from the encoder, and the cluster labels from the DICOM tags, we calculated the mean cosine distance between images and cluster centres, and compared it against the mean cosine distance of visual features obtained using randomly permuted cluster labels. For the DICOM tags, the mean distance between objects having visual features is 0.28 and for randomly permuted cluster labels, the mean distance is 0.42. Standard deviation of the mean distances is 0.023 and 0.017, respectively. For testing, paired t-test was used. The null hypothesis, stating that there is no difference in the mean distance from the cluster centres between visual features with cluster labels inferred by clustering DICOM tags and randomly permuted cluster labels, is rejected for $t(9), p < 0.001$, the test statistic being 16.42. This confirmed that the DICOM tags also cluster objects with respect to their visual similarity (in the pixel space).

4 CONCLUSION

The evaluation procedure described in this paper shows the information potential of DICOM tags for grouping related images into visually similar groups. We prove that DICOM-based clustering can provide a useful input towards assembling visually more homogeneous clusters of images, when compared against randomly grouped images. We estimated the visual similarity of images using an unsupervised embedding of the domain images.

Although the results in this insight study look promising, further work on this topic is necessary. There are many directions in which the research can be expanded. Clustering performance can be improved by using different feature selection algorithms. Various dimensionality reduction algorithms can also be considered because they reduce noise (unwanted variance) and offer the possibility for intuitive data visualisation. Furthermore, developing additional
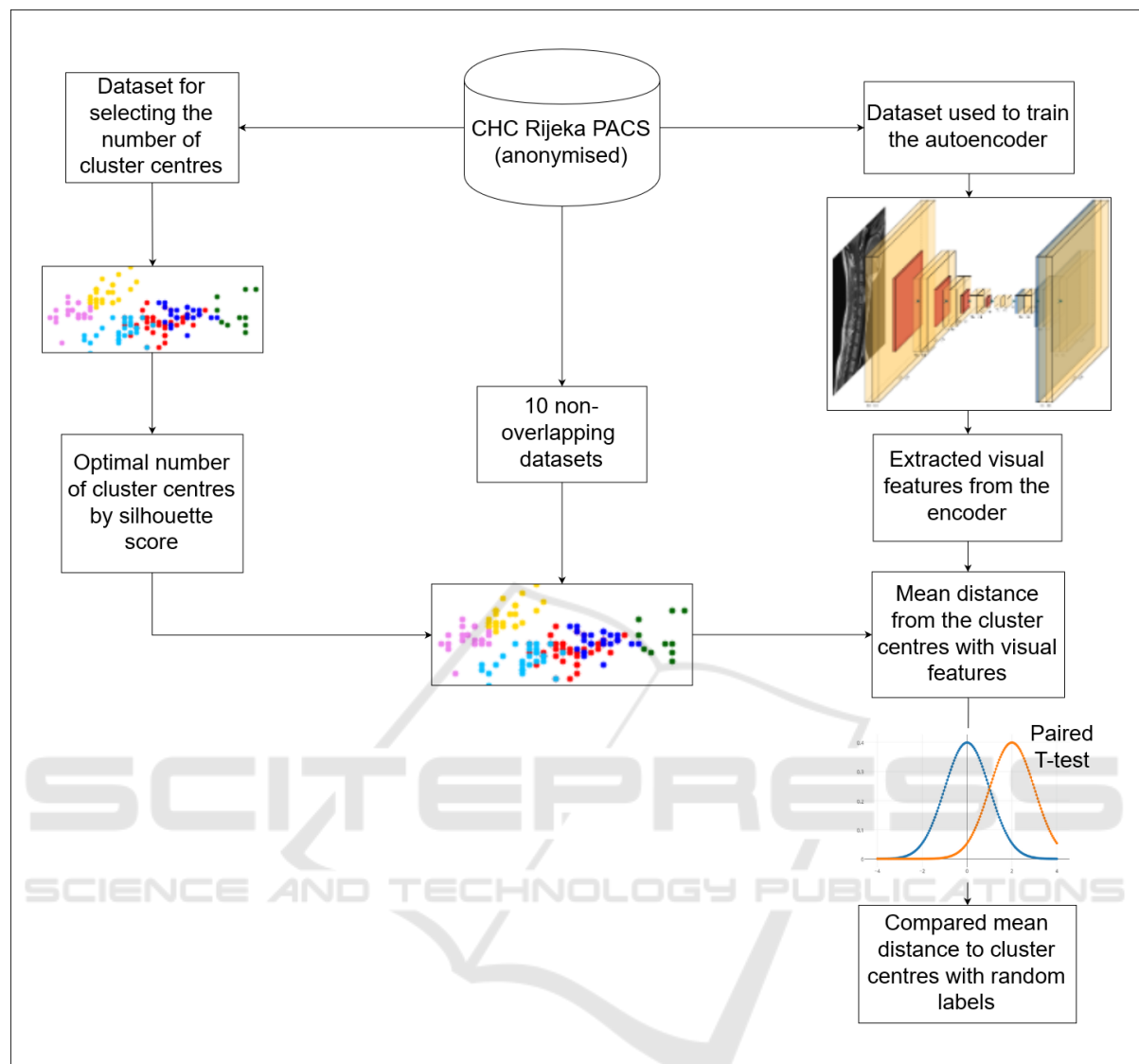
Figure 2: A graphic depiction of the experimental setup, specifically concerning the evaluation procedure used. All mentioned datasets are mutually disjoint (i.e. non-overlapping).

manually-tailored data parsing techniques for exploring DICOM tags should also be explored. Clustering performance can probably be improved in addition by using different clustering algorithms. This would include the evaluation of the existing model and comparing it against other techniques, such as the agglomerative hierarchical clustering (Day and Edelsbrunner, 1984) which might provide even better clustering results. In addition, merging different repositories of medical images could prove useful because different medical repositories contain images with tag-assignment standards that could differ significantly, which is an additional challenge. Another application involving the use of DICOM data is related to impu-

tation and correcting of missing and falsely inserted tags, because DICOM tags can be prone to human errors. These ideas can be used to improve the quality of the set foundation for deriving the semantic structure of a medical radiology dataset.
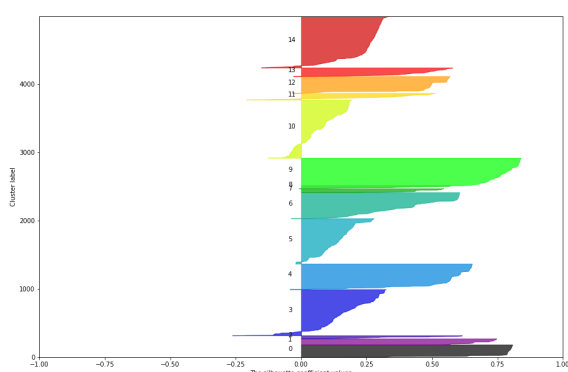
ACKNOWLEDGEMENTS

Figure 3: *Silhouette score* for 5000 data instances, using the number of clusters $K = 15$.

REFERENCES

Avishkar Misra, Mamatha Rudrapatna, and Arcot Sowmya (2005). Automatic Lung Segmentation: A Comparison of Anatomical and Machine Learning Approaches. In *Proceedings of the 2004 Intelligent Sensors, Sensor Networks and Information Processing Conference*.

Bidgood, W. D., Horii, S. C., Prior, F. W., and Van Syckle, D. E. (1997). Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging. *Journal of the American Medical Informatics Association*, 4(3):199–212.

Chen, M., Shi, X., Zhang, Y., Wu, D., and Guizani, M. (2017). Deep Features Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network. *IEEE Transactions on Big Data*.

Choplin, R. H., Boehme, J. M., and Maynard, C. D. (1992). Picture archiving and communication systems: an overview. *RadioGraphics*, 12(1):127–129.

Day, W. H. E. and Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1):7–24.

Dimitrovski, I., Kocev, D., Loskovska, S., and Džeroski, S. (2011). Hierarchical annotation of medical images. *Pattern Recognition*, 44(10-11):2436–2449.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*.

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*, volume 4685, pages 280–287. SPIE.

Källman, H. E., Halsius, E., Olsson, M., and Stenström, M. (2009). DICOM metadata repository for technical information in digital medical images. *Acta Oncologica*, 48(2):285–288.

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *ImageNet Classification with Deep Convolutional Neural Networks*, pages 1097–1105. Curran Associates, Inc.

L., K. and P., R. (1987). Clustering by means of Medoids. In *Statistical Data Analysis Based on the L1 Norm and Related Methods*.

Lehmann, T. M., Schubert, H., Keysers, D., Kohnen, M., and Wein, B. B. (2003). The IRMA code for unique classification of medical images. In *Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*, volume 5033, page 440. SPIE.

Masci, J., Meier, U., Cireşan, D., and Schmidhuber, J. (2011). Stacked convolutional auto-encoders for hierarchical feature extraction. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

Park, H. S. and Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*.

Qiang Yang and Pan, S. J. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

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.

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

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. cite arxiv:1409.1556.

Van Soest, J., Lustberg, T., Grittner, D., Marshall, M. S., Persoon, L., Nijsten, B., Feltens, P., and Dekker, A. (2014). Towards a semantic PACS: Using Semantic Web technology to represent imaging data. *Studies in health technology and informatics*, 205:166–70.

Xie, J., Girshick, R., and Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. In *33rd International Conference on Machine Learning, ICML 2016*.

Yang, Y., Xu, D., Nie, F., Yan, S., and Zhuang, Y. (2010). Image clustering using local discriminant models and global integration. *IEEE Transactions on Image Processing*.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc.

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