Proc. of the International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)
16-18 November 2022, Maldives

2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) | 978-1-6654-7095-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICECCME55909.2022.9988310

Using Autoencoders to Reduce Dimensionality of DICOM Metadata

Mateja Napravnik
University of Rijeka, Faculty of Engineering
Vukovarska 58, Rijeka 51000, Croatia,
mnapravnik@riteh.hr

Franko Hržić
University of Rijeka, Faculty of Engineering
Vukovarska 58, Rijeka 51000, Croatia,
fhrzic@riteh.hr

Robert Baždarić
University of Rijeka, Faculty of Engineering
Vukovarska 58, Rijeka 51000, Croatia,
rbazdaric@riteh.hr

Sebastian Tschauner
Medical University of Graz
Auenbruggerplatz 34, Graz 8036, Austria
sebastian.tschauner@medunigraz.at

Damir Miletić
University of Rijeka, Clinical Hospital Centre Rijeka,
Krešimirova 42, Rijeka 51000, Croatia
damir.miletic@medri.uniri.hr

Mihaela Mamula
University of Rijeka, Clinical Hospital Centre Rijeka,
Krešimirova 42, Rijeka 51000, Croatia
mihaela.mamula@gmail.com

Ivan Štajduhar
University of Rijeka, Faculty of Engineering
Vukovarska 58, Rijeka 51000, Croatia,
istajduh@riteh.hr

***Abstract*—Digital Imaging and Communication in Medicine (DICOM) is a standardized format for storing medical images enriched with descriptive data. It consists of a number of informative header tags. E.g., these tags contain information concerning imaging technique, patient weight, patient age and so on. In machine learning, these tags can, among other things, be used for categorization, classification, and general manipulation of DICOM data. However, because of their number and variety, using them for automation undoubtedly impacts computational performance. To tackle this problem, the possibility of using autoencoders is explored in this manuscript. It was hypothesized that clustering data compressed with autoencoders can achieve the same results as when working with raw – uncompressed data. To support this claim, clustering of compressed and uncompressed data was performed on a dataset of $25,000$ DICOM files from the Clinical Hospital Centre Rijeka PACS. The results show no significant difference between compressed and uncompressed data ($p > 0.05$), thus confirming that equally good clustering results can be accomplished using a smaller representation.**

***Index Terms*—DICOM, Autoencoders, Medical Imaging, Clustering**

I. Introduction

Over the past few decades, technological advances called for more medical imaging standardization. Picture Archiving and Communication Systems (PACS) [1] were put in place to allow more accessible storage, manipulation, and retrieval of medical images. Different file formats for medical images were developed [2], notably the Digital Imaging and Communication in Medicine (DICOM) format. The DICOM format consists of two parts: the image (the raw pixel data) and corresponding metadata located in the file's header [3]. The metadata can be described as a collection of structured tags which were either manually set by medical professionals or automatically inputted by imaging devices. Images can be obtained through different imaging techniques such as Computed Tomography (CT), Computed Radiography (CR), Magnetic Resonance (MR), and so on.

Analysis of large medical repositories can be challenging [4]. Manual filtering and noise removal on large datasets is a difficult task to complete, which is why different approaches to automating the process are being explored [5]. One of these approaches is clustering, which can also be used to label unstructured medical data or help in computer-aided diagnosis (CAD) [6]. However, as the dataset size increases, so does the amount of resources needed to complete clustering tasks. To tackle this, clustering can be sped up by using dimensionality reduction methods [7].

The DICOM metadata consists of both numerical and categorical data. Raw categorical data are generally unsuitable to serve as algorithm inputs; hence they are often encoded as one-hot vectors [8], which is where problems arise for high-cardinality categorical variables. The increase in dimen-

978-1-6654-7095-7/22/$31.00 © 2022 European Union

sionality caused by such variables can impact performance. Therefore, one can opt to reduce the number of features by using autoencoders (AEs) [9].

AEs learn to map a set of data into a latent-space representation having fewer dimensions than the original data [10]. As such, the latent-space representation can be considered a compressed version of the input. Yildirim et al. [11] showcased how a convolutional autoencoder (CAE) could reduce dimensionality of electrocardiograms (ECG) to allow for easier data transfer. Besides being considered a compression method, AEs can also be used as feature extractors. The latent-space representation can be fed as input into other algorithms and networks. Guan et al. [12] describe using an AE as a feature extractor for human embryonic stem cell videos, whose latent-space output is then forwarded to a RandNet neural network for classification. Their approach outperformed other state-of-the-art methods, with a classification accuracy of $97.23 \pm 0.94\%$. In addition, it has been shown that CAEs can be utilized to extract encoded representations of images which can later be used to compare the visual similarity of clustered groups [13].

Research utilizing DICOM tags is scarce but not entirely unexplored. Several papers have described the usage of DICOM headers to influence or manipulate the respective pixel data. Guld et al. [14] stated it would be impossible to achieve automated categorization solely through DICOM tags. However, they focused only on the *BodyPartExamined* tag and its effectiveness (or rather, the lack thereof) to accurately describe the examined anatomic region. While it is true that manually inputted tags are more susceptible to noise, this does not mean DICOM metadata should be excluded outright. Kim et al. [15] have shown that the fidelity of JPEG compressed CT images can be predicted solely using DICOM metadata. Furthermore, Manojlović et al. [16], [17] have shown that generating pairwise constraints from DICOM metadata can improve the results of medical image clustering.

This manuscript explores the possibility of reducing the dimensionality of DICOM tags through the usage of AEs. It gives a comparison of the performance of clustered latent-space representations to clusters obtained from the original data. Finally, an analysis whether the encoded metadata can be used to group DICOM images into visually similar groups is made. To the best of our knowledge, no similar research has been done concerning DICOM metadata compression.

II. Materials and Methods

This section describes the origin of data and its preprocessing for utilization by AE networks and clustering algorithms. Furthermore, it reports on how the evaluation was performed and which tests were used. The general pipeline of the process described in this section is shown in Fig. 1.

A. Dataset

The dataset was obtained from the PACS system used in Clinical Hospital Centre (CHC) Rijeka. It consists of approximately 30 million DICOM images [4] gathered through

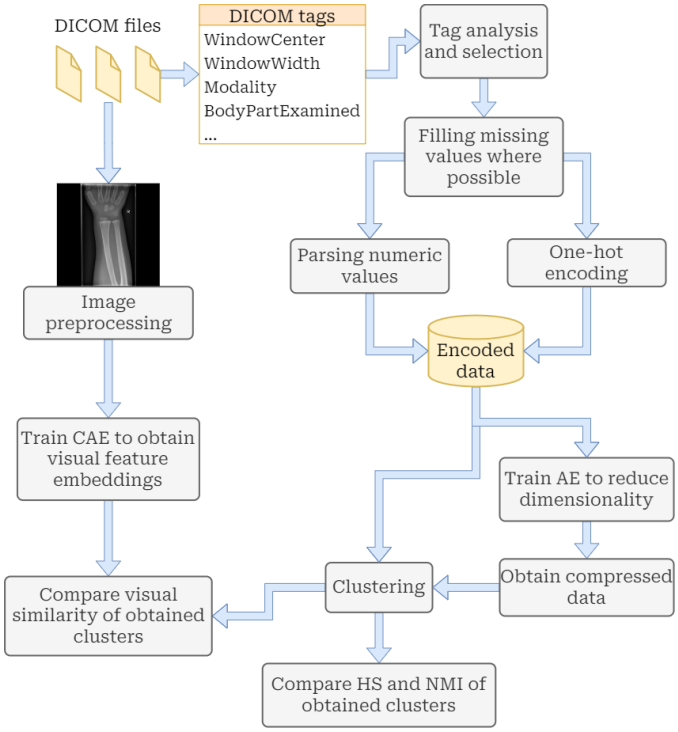

Fig. 1. The general experimental pipeline. The DICOM tags were analysed and prepared for clustering, while the images were preprocessed and then used to compare visual similarities of obtained clusters.

standard clinical practice between 2010 and 2017. From these images, approximately $25,000$ images were subsampled in such a was that they met the following criteria: the pixel data was grayscale, the DICOM tags *WindowCenter* and *WindowWidth* were set, and the image had been captured in one of the six chosen modalities – CT, CR, MR, X-ray Angiography (XA), Nuclear Medicine (NM) and Radio Fluoroscopy (RF). The subset was balanced with respect to modality, with each modality having approximately $4,000$ instances in the chosen dataset subsample. In this subsample, a total of 28 different distinct *BodyPartExamined* values were recorded. Unlike *Modality*, balancing the dataset with regard to *BodyPartExamined* proved to be difficult, due to different modalities being suitable for observing and emphasizing different parts of a body [18]. The dataset was divided into three parts: the training subset consisted of approximately $18,000$ instances, the validation subset counted $2,000$ instances, and the test subset consisted of $5,000$ instances of DICOM files.

1) DICOM Tags Analysis and Preprocessing: A total of 651 different tags occurred at least once in the subset. However, many of these proved to be uninformative, having less than two unique values in the entire subset. By omitting these tags, the number of tags was reduced to 424. Further inspection of the tags revealed that a lot of them were scarcely used, which hindered their ability to be useful. A threshold of 20% was chosen, meaning that any tag that had been used in less than 20% of the subset was also dropped. The reduction resulted in the total number of tags being 76. The remainder

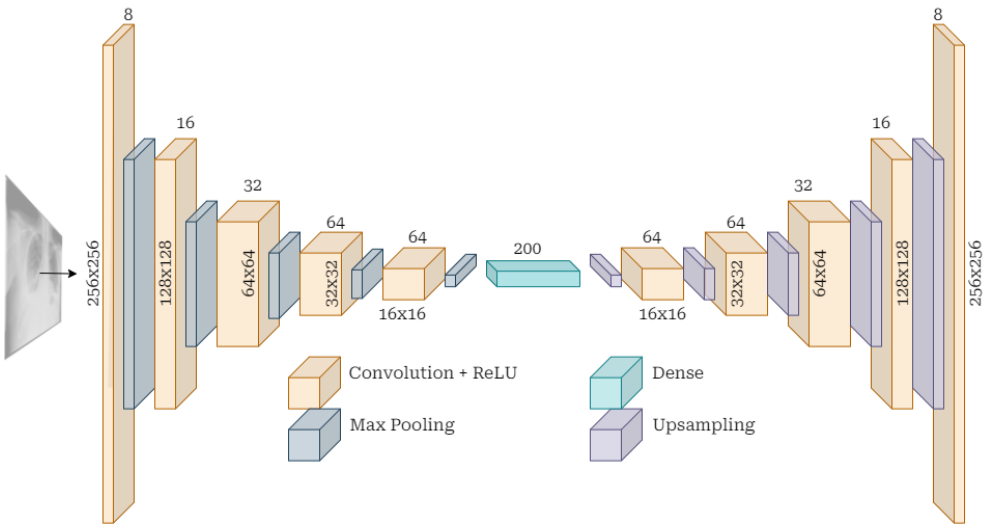

Fig. 2. Architecture of the convolutional autoencoder used to generate visual feature embeddings.

was inspected manually which led to the removal of tags such as *StudyDescription*, *ProtocolName* and *ImageComments*. Although informative in their own right, these tags contain natural language that requires processing that falls outside the scope of this paper. As a result, the tag count dropped to 68. Although these tags were removed from the dataset, they still proved useful in filling missing values (through regular expressions) for other tags, namely *BodyPartExamined*.

Some of the tags, such as *AcquisitionMatrix* or *PixelSpacing* can contain two or more numeric values. These types of tags were then split into multiple values; for example, *PixelSpacing* was transformed into *PixelSpacing0* and *PixelSpacing1*. This way, the feature count was increased from 68 to 77. The numeric values were scaled to fit the range $[0, 1]$, while the categorical tags were one-hot encoded. The final encoded dataset contained 565 columns.

2) Image Preprocessing: One of the prerequisites mentioned above when choosing which DICOM files to use was that the file should have *WindowCenter* and *WindowWidth* contained within the header. Window center and window width are used to accentuate or suppress different parts of the image [19]. If the image contained more than one window center or width, the first value was used to window the image. DICOM images usually have a pixel depth between 12 or 16 bits [2]. After windowing, the images were scaled down to 8-bit integers and resized to 256×256 pixels using bilinear interpolation. To preserve the aspect ratio, zero-padding was added where necessary.

B. Autoencoder Network

Multiple AE networks were trained, with each differing in depth, layer sizes, bottleneck sizes, and learning rates. All models were trained across 10 epochs with a batch size of 64. Ultimately, the model with the lowest reconstruction error, while still having a satisfactory compression rate, was chosen. The encoder comprised of four dense layers of sizes 565, 256, 128 and 64, respectively. Each layer was followed by a batch normalization layer and a leaky rectified linear unit (ReLU) activation function [20]. The bottleneck layer was capped at 56 features (10% of the original size). Next, it was followed by the decoder part of the network, which mirrored the layout of the encoder. Sigmoid was the chosen activation function in the network output. Adam was used as the optimizer, while mean squared error (MSE) was used as the loss function. After training, the encoder would be used to transform data into a latent-space representation. Henceforth, the data from latent-space of the encoder will be referred to as *compressed data*, while the uncompressed data will be denoted as *original data*.

C. Evaluation

The evaluation was two-fold. First, the original and compressed data were clustered using K-medoids, as shown in Fig. 1. K-medoids is a clustering algorithm where cluster centers (medoids) are chosen from the input data to minimize the dissimilarity of all dataset points to the nearest medoid. Since the medoids are selected from the input data, K-medoids are less susceptible to outliers [21]. The Manhattan distance was used as the clustering distance metric. Then, a comparison was made to see if, when clustered, the compressed data would be divided into visually similar groups. To test this, the same approach used in [13] was performed.

To compare the visual similarity of resulting clusters, a latent-space representation of images was calculated using a CAE model. Images (which were previously windowed and converted into an 8-bit representation) were normalized to fall in the range $[0, 1]$. Multiple CAE architectures with differing learning rates were tested, and, ultimately, the one obtaining the lowest reconstruction error while still having an adequate bottleneck size was chosen. The CAE model consisted of two parts: encoder and decoder. The encoder consisted of five convolutional layers of 3×3 kernel size, followed by a 2×2 max-pooling layer. The first three convolutional layers had 8,

16, and 32 filters, respectively; while the latter two comprised 64 filters each. The bottleneck layer was a dense layer of size 200. The decoder network follows a mirrored layout of the encoder. The full architecture of the model is shown in Fig. 2. Adam was used as the optimizer, with a learning rate of 10^{-6}. MSE was chosen as the loss function. The model was trained across 100 epochs, with a batch size of 32.

Once a dataset had been clustered, a comparison of visual similarity between clusters was conducted. Since it had been proven that DICOM tags could group images that are visually similar [13], the visual similarity of groups clustered with compressed data was compared against the visual similarity of groups clustered with original data. The comparison was conducted by calculating the cosine distance from each image to its nearest medoid. Given visual feature embedding x and its nearest medoid y, their cosine dissimilarity is calculated as follows:

$$d(x, y) = 1 - \frac{x \cdot y^T}{\|x\| \cdot \|y\|} \tag{1}$$

A test was performed to see whether the original and compressed data differ in inhomogeneity and mutual information when clustered. It can be expected that the original data will be homogeneous regarding imaging modality and body part examined. A comparison was made to see if the same rule also follows compressed data. To test this, the homogeneity score (HS) and normalized mutual information (NMI) with regard to *Modality* and *BodyPartExamined* were calculated, for both compressed and original data. HS and NMI are expressed in the range $[0, 1]$, where the highest score of 1.00 indicates that the data is homogeneous, meaning that all members of a specific class are assigned to the same cluster. On the other hand, a score of 0.00 indicates no homogeneity with respect to a class within clusters. NMI is calculated as:

$$NMI(l_T, l_P) = \frac{I(l_T, l_P)}{H(l_T) \cdot H(l_P)}, \tag{2}$$

where l_T are the ground truth labels, l_P are the predicted cluster labels, $I(l_T, l_P)$ is the mutual information between the two, and $H(l_T)$ and $H(l_P)$ denote the entropy of both variables. Homogeneity, on the other hand, is calculated as:

$$h = 1 - \frac{H(C|K)}{H(C)}, \tag{3}$$

where C are the classes, K are the cluster assignments, $H(C|K)$ is their conditional entropy, and $H(C)$ is the entropy of classes.

After calculating NMI and HS scores for different numbers of clusters, the results were compared using statistical tests.

III. Results and Discussion

Different cluster counts were tested, $K \in \{4, 7, 11, 15, 20, 25, 30, 40, 50, 75, 100, 150\}$. NMI and HS were computed with regard to body part examined (BPE) and modality (Mod). This adds up to four different metrics (HS Mod, HS BPE, NMI Mod, NMI BPE) and each of them was calculated for every cluster count K.

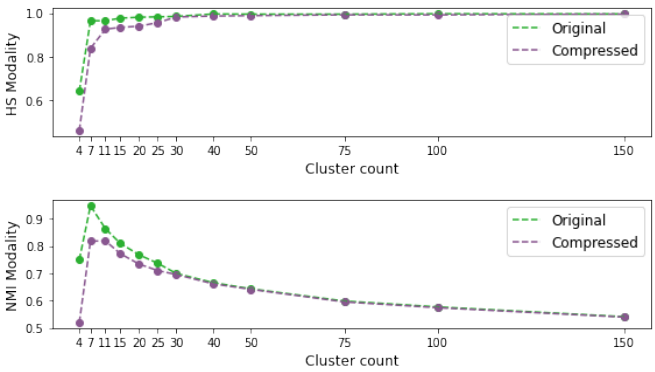

Fig. 3. Homogeneity and normalized mutual information of clusters with regard to modality.

HS and NMI regarding modality are shown in Fig. 3. The compressed data has slightly worse homogeneity score for $K \in \{4, 7\}$ than the original data, but it improves for $K \in \{11, 15, 20\}$. For $K \geq 25$, the difference is almost imperceptible. The very same can be said about NMI regarding modality. This means that, for a small number of clusters, the homogeneity and mutual information slightly differ, yet the difference is almost indiscernible for $K \geq 25$.

HS and NMI regarding body part examined are shown in Fig. 4. The compressed data's homogeneity is almost exact as the original data, for all cluster counts K. On the other hand, NMI regarding body part examined fluctuates. Compressed data performs slightly worse than original data for smaller cluster counts $K \in \{4, 7, 11, 15, 20\}$, whereas for $K \in \{30, 50, 75, 100\}$ it marginally outperforms the original data.

While HS with regard to modality ranges high ($\langle 0.45, 1.00 \rangle$) for different K, HS with regard to body part examined ranges slightly lower ($\langle 0.22, 0.85 \rangle$). This can be attributed to *BodyPartExamined* having higher cardinality than *Modality*. As was mentioned before, the data contained 28 distinct values for

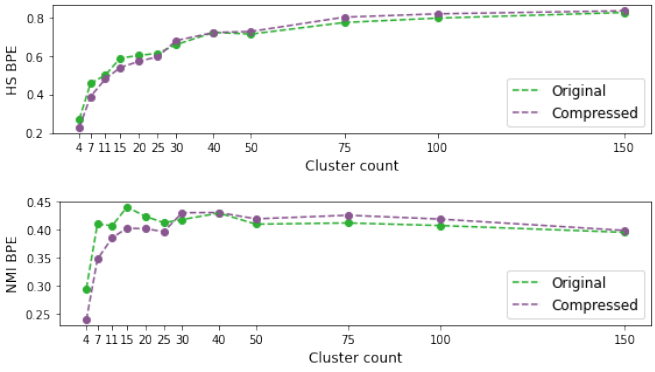

Fig. 4. Homogeneity and normalized mutual information of clusters with regard to body part examined.

body part examined, yet there were only 6 different modalities. Nonetheless, the obtained clusters achieved a satisfactory level of homogeneity and mutual information with regard to both variables.

Before comparing the four metrics, each was subjected to a Shapiro-Wilk test of normality. If the Shapiro-Wilk test showed the results were normally distributed, then they could be compared using a two-tailed paired T-test. Otherwise, they were subjected to a two-tailed Mann-Whitney U-test. The results are depicted in Table I. The table shows no significant differences between compressed and original data ($\alpha = 0.05$). This means the compressed data, when clustered, gives equally good results as the original data, despite being only 10% of its size.

TABLE I.
COMPARISON BETWEEN NMI AND HS OF COMPRESSED AND ORIGINAL DATA.

Metric	Shapiro-Wilk p-value	Test performed	p-value
HS Mod	Original: $5.4 \cdot 10^{-6}$ Compressed: $5.1 \cdot 10^{-5}$	Mann-Whitney U	0.157
NMI Mod	Original: 0.978 Compressed: 0.673	T-test	0.053
HS BPE	Original: 0.466 Compressed: 0.461	T-test	0.24
NMI BPE	Original: 0.0002 Compressed: 0.0007	Mann-Whitney U	0.507

The elbow method was used to find the optimal number of clusters. The observed metric for finding the elbow was the sum of squared distances, shown in Fig. 5. Both the original and compressed data exhibited an elbow around $K = 15$. Hence, this cluster count was used to compare the visual similarity of images within clusters. The average cosine distance within clusters was 0.0644 ± 0.063 for compressed data, while it was 0.0661 ± 0.0662 for original data. The cosine distances did not follow the normal distribution (Shapiro-Wilk; $p_{\text{compressed}} = 0.00$, $p_{\text{original}} = 0.00$, $\alpha = 0.05$). A two-tailed Mann-Whitney U-test was performed, giving a p-value of $p = 0.945$, $\alpha = 0.05$. There is no statistically significant difference in visual similarities between compressed and original data. Therefore, the compressed data can produce groups that are visually similar.

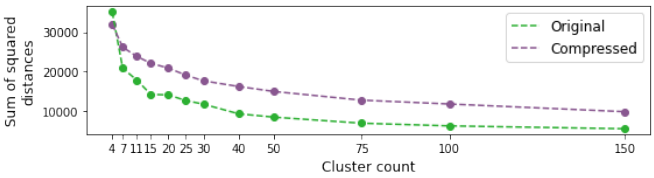

Fig. 5. Sum of squared distances. Used for the elbow method.

IV. CONCLUSION

In this work, the possibility of reducing the dimensionality of DICOM tags was presented with promising results. This manuscript has proven that, when clustered, compressed data can still be homogeneous regarding modality and BPE. The compressed data had a 10 times smaller representation than the original data, yet there was no statistically significant difference between the homogeneity and mutual information of obtained clusters. Furthermore, it has been shown that the compressed data can be clustered into groups whose members share a visual similarity, akin to how original data can be clustered into visually similar groups. Although this research was completed on a relatively small subset of the whole dataset ($25,000$ versus 30 million in available DICOM files), it offers insight into how DICOM tags can be compressed and still retain enough information. This means that equally good clustering results can be achieved using a compressed representation, which can speed up performance of grouping a large medical repository, such as the one obtained from CHC Rijeka PACS.

Although this research delivers promising results, the topic could be more delved into. Aside from AEs, other dimensionality reduction methods could cope with the presented task. Moreover, some of the DICOM tags were removed from the dataset due to their content, which includes natural language. These tags contain possibly helpful information, and their processing, as well as testing other dimensionality reduction methods, could be researched further.

V. ACKNOWLEDGMENT

This work has been supported in part by Croatian Science Foundation [grant number IP-2020-02-3770]; and by the University of Rijeka, Croatia [grant number uniri-tehnic-18-15].

REFERENCES

[1] N. Strickland, "PACS (picture archiving and communication systems): filmless radiology," *Archives of disease in childhood*, vol. 83, p. 82–86, 2000.

[2] M. Larobina and L. Murino, "Medical Image File Formats," *Journal of Digital Imaging*, vol. 27, pp. 200–206, APR 2014.

[3] M. Mustra, K. Delac, and M. Grgic, "Overview of the DICOM standard," vol. 1, pp. 39 – 44, 10 2008.

[4] I. Štajduhar, T. Manojlović, F. Hržić, M. Napravnik, G. Glavaš, M. Milanič, S. Tschauner, M. Mamula Saračević, and D. Miletić, "Analysing Large Repositories of Medical Images," in *International Conference on Bioengineering and Biomedical Signal and Image Processing*, pp. 179–193, Springer, 2021.

[5] M. E. Tschuchnig and M. Gadermayr, "Anomaly Detection in Medical Imaging-A Mini Review," *Data Science–Analytics and Applications*, pp. 33–38, 2022.

[6] A. Garg and V. Mago, "Role of machine learning in medical research: A survey," *Computer Science Review*, vol. 40, p. 100370, 2021.

[7] Y. Hozumi, R. Wang, C. Yin, and G.-W. Wei, "UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets," *Computers in biology and medicine*, vol. 131, p. 104264, 2021.

[8] K. Potdar, T. Pardawala, and C. Pai, "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," *International Journal of Computer Applications*, vol. 175, pp. 7–9, 10 2017.

[9] S. Mumtaz and M. Giese, "Hierarchy-based semantic embeddings for single-valued & multi-valued categorical variables," *Journal of Intelligent Information Systems*, vol. 58, pp. 613–640, JUN 2022.

[10] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1798–1828, AUG 2013.

[11] O. Yildirim, R. San Tan, and U. R. Acharya, "An efficient compression of ECG signals using deep convolutional autoencoders," *Cognitive Systems Research*, vol. 52, pp. 198–211, DEC 2018.

[12] B. X. Guan, B. Bhanu, R. Theagarajan, H. Liu, P. Talbot, and N. Weng, "Human embryonic stem cell classification: random network with autoencoded feature extractor," *Journal of Biomedical Optics*, vol. 26, MAY 2021.

[13] T. Manojlovic, D. Ilic, D. Miletic, and I. Stajduhar, "Using DICOM Tags for Clustering Medical Radiology Images into Visually Similar Groups," in *ICPRAM: Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods*, pp. 510–517, 2020.

[14] M. Guld, M. Kohnen, D. Keysers, H. Schubert, B. Wein, and T. Lehmann, "Quality of DICOM header information for image categorization," *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 4685, 03 2002.

[15] K. J. Kim, B. Kim, H. Lee, and H. Choi, "Predicting the fidelity of JPEG2000 compressed CT images using DICOM header information," *Medical Physics*, vol. 38, pp. 6449–6457, DEC 2011.

[16] T. Manojlović and I. Štajduhar, "Deep Semi-Supervised Algorithm for Learning Cluster-Oriented Representations of Medical Images Using Partially Observable DICOM Tags and Images," *Diagnostics*, vol. 11, no. 10, 2021.

[17] T. Manojlović, M. Milanič, and I. Štajduhar, "Deep embedded clustering algorithm for clustering PACS repositories," in *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 401–406, IEEE, 2021.

[18] D. Bell, " Modality ." https://radiopaedia.org/articles/61013. Accessed on 22 Aug 2022.

[19] A. Murphy and Y. Baba, " Windowing (CT)." https://radiopaedia.org/articles/52108. Accessed on 22 Aug 2022.

[20] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," *CoRR*, vol. abs/1803.08375, 2018.

[21] L. Rdusseeun and P. Kaufman, "Clustering by means of medoids," in *Proceedings of the statistical data analysis based on the L1 norm conference, Neuchatel, Switzerland*, vol. 31, 1987.