Marek SOCHA[1], Aleksandra SUWALSKA[1], Wojciech PRAZUCH[1],
Michał MARCZYK[1], Joanna POLANSKA[1], POLCOVID Study Group

Chapter 9. UMAP-BASED GRAPHIC REPRESENTATION OF POLCOVID CHEST X-RAY DATA SET HETEROGENEITY

9.1. Introduction

When analysing real problems, on real data, the researchers want to have a great understanding of underlying structures and existing patterns, which usually requires use of sophisticated and accurate visualisation techniques. In many fields, real datasets often have high dimensionality, which makes the computations complicated and time-consuming. Visualisations of high-dimensional data is difficult or even impossible in some cases. One way to overcome this issues is dimensionality reduction by projection of a data into low-dimensional space. However, most of the techniques are mainly dedicated to tabular data, making the projections of images less reliable.

Uniform Manifold Approximation and Projection (UMAP) is an embedding technique that projects high-dimensional data points into low-dimensional space [1]. It is a useful tool for data visualisation and pre-processing, with the potential to be used as a clustering method. Despite being mainly used for tabular data, some attempts have been made to apply UMAP to image data analysis.

The first usage of the UMAP embedding in the scope of image data analysis was described in the original paper [1]. Images was firstly vectorized and then passed to UMAP procedure. The experiments were made using the Pen digits [2], COIL 20 [3] and COIL 100 [4] datasets. This method was also used in other articles [5–7] which operated on MNIST [8], Fashion-MNIST [5], USPS [9] and Afro-MNIST [10] datasets.

[1] Department of Data Science and Engineering, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland. Corresponding author: aleksandra.suwalska@polsl.pl.

(a) COIL-20 (b) COIL-100 (c) MNIST

Fig. 9.1. Example images from commonly used benchmark datasets (a) COIL-20, (b) COIL-100, (c) MNIST

Rys. 9.1. Przykładowe obrazy z powszechnie używanych zestawów danych porównawczych (a) COIL-20, (b) COIL-100, (c) zbiór MNIST

These datasets consist of clearly defined, low-resolution images with a neutral background, either black or white. Image shapes are adjusted so the Region Of Interest (ROI) is dominant. However, these properties are not commonly seen in all real world image data sets, where the object of interest could be present in different locations across set of images, could have varying sizes and could be displayed with diverse backgrounds.

The ability to generate embeddings that will separate different categories of images in the 2D plane could be very helpful in medical image analysis to reveal differences and similarities between the analysed groups. There are examples of UMAP usage in this context in articles [11–13]. Each of them, aimed to describe structures of a Tumours and Nodules in Computed Tomography scan (CT), Positron Emission Tomography (PET) or Magnetic Resonance imaging (MRI) images. In those cases, ROI is located exactly on the area of interest. It is different in analysis of chest X-Ray (CXR) images, where we observe different shapes, alignment, intensities, background values, artefacts but more importantly, ROI placement. The place in which a disease that define lung change occurs is heterogeneous in shape and placement. Thus, it is very hard to precisely indicate it in the CXR image. Forcing the use of whole or a part of CXR image region, where in such a case the UMAP procedure could potentially capture information that is not related to the biological differences observed in a lung region.

The goal of the study was to obtain visual separability of analysed disease entities on UMAP plots in order to make this information useful for further analysis. This study proposes a novel method of semi-supervised UMAP embedding creation, suitable for the CXR image analyses. The method is robust to the ROI placement, image size, image alignment and type of background.

9.2. Materials and Methods

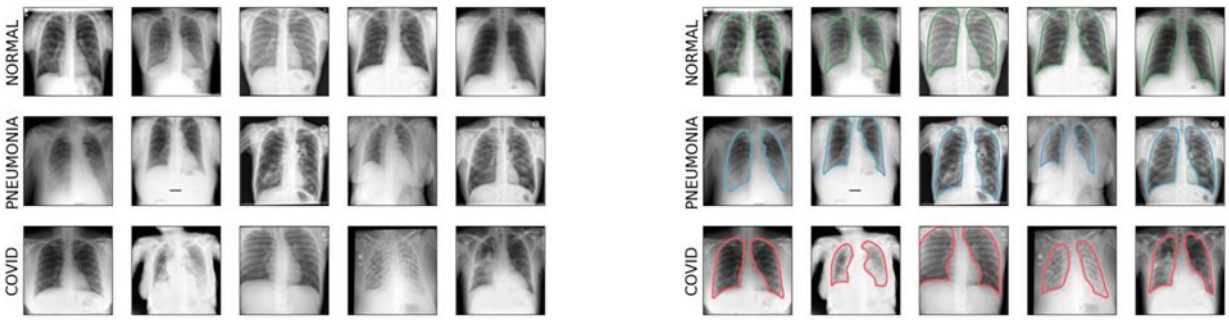

Fig. 9.2. POLCOVID database example images. (a) Example images from POLCOVID database, (b) Example images from POLCOVID database with segmentations
Rys. 9.2. (a) Przykładowe dane z bazy POLCOVID, (b) Przykładowe dane z bazy POLCOVID wraz z segmentacjami

In the study, images from the POLCOVID database were used. POLCOVID is an original and unique COVID-19 database containing X-Ray images. A data incorporated within the database were collected from Polish hospitals in the CIRCA project (COVID-19 RTG/CT-Based Diagnosis). The considered subset of POLCOVID database data consisted of 4956 CXR images, 2578 of which were normal (healthy), 1174 were non-COVID pneumonia and 1204 were COVID-19 (with positive RT-PCR test result). The images were collected from 24 different clinical hospitals in Poland, so they were acquired with different RTG devices and with different parameters of image scanning. Therefore, the dataset is heterogeneous in terms of pixel intensity and image resolution.

9.2.1. Image pre-processing

The images had to be properly standardised, in order to minimise the batch effect of obtaining data from various sources. The presence of white artefacts, like text on image, has a significant impact on the pixel intensity distribution of an image. To overcome this problem, 0.25% of extreme pixel intensity values were removed from images before the standardisation to the range of [0, 1]. This step made the image distributions closely related. All images were resized to 512x512 pixels.

9.2.2. Radiomic Features

Radiomic Features [14–17] are a large set of numerical values aimed towards describing medical image fragment indicated by a given ROI. The usage of radiomic features was proven to be effective in task of cancer detection and evaluation [11–13] both in terms of UMAP visualisation and diagnostic potential. The extraction of the radiomic features was made using python package pyradiomics [18].

9.2.3. Features preprocessing

In the method's pipeline, in the first step, from the set of whole features those with variance equal to zero were filtered out. Remaining features were scaled to the range of [0,1]. For the initial dimension reduction, the PCA technique [19] were used from which the features which explained from 85% to 95% variance were taken. This percentage value was adjusted depending on the final amount of PCA components.

9.2.4. U-Net segmentation

U-Net [20] is a neural network architecture for image semantic segmentation. For the purpose of lung region segmentation some hyper-parameters and features of the model were adjusted. On the course of experiments, the final U-Net based model consisted of 4 encoding and decoding layers with SELU activation function [21], Batch Normalisation and Dropout with drop out coefficient of 0.5. The number of filters for each convolutional layer ranged from 32 to 512 depending on depth of the convolutional layer. The input images were resized and padded to the shape of 512x512. Neural network were trained using over 5200 train images with hand crafted lung masks.

9.2.5. Effect size

Effect size is a quantitative measure of the strength, of a phenomenon calculated on the basis of data [22], in contrast to p-value which indicates whether an effect exists, not how great it is.

9.2.6. Features selection

Features selection process was based on the effect size. The non-parametric Kruskal-Wallis test [23] was performed on the features calculated for each group of patients and for each feature an effect size was computed. Then, the features were ordered according to the effect size (Figure 9.7) as a measure of feature importance in the task of distinguishing the patient categories. Final number of selected features were decided on the course of experiments described in the Results in subsection 3.4.

9.3. Results

9.3.1. UMAP on whole images

The most common approach for applying UMAP on images is flattening the images into vectors and storing them in a tabular manner. This method was successful while dealing with the distinction between classes on datasets like Fashion-MNIST [5] or COIL-100 [1, 4]. The result of the standard UMAP approach applied to the POLCOVID CXR dataset is shown on the scatter plot in Figure 9.3. The hyperparameters used for UMAP are as follows: metric=euclidean, n_neighbors=500 and min_dist=0.8.

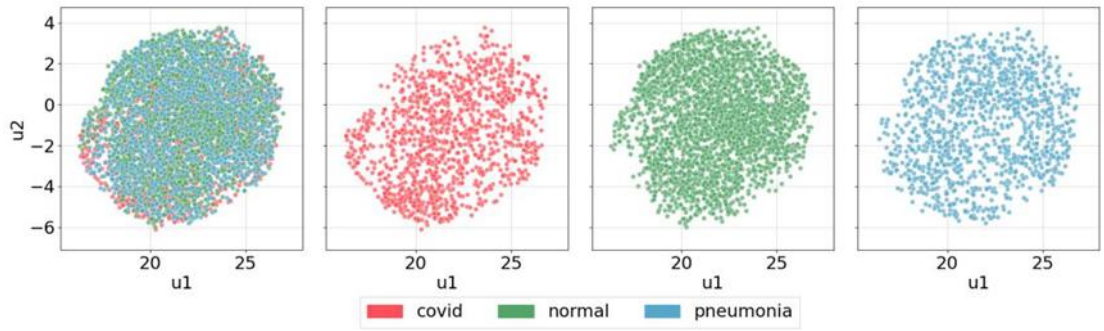

Fig. 9.3. Result of the UMAP procedure on the vectorized images
Rys. 9.3. Rezultat techniki UMAP wyliczonej na obrazach zwektoryzowanych

The scatter plot, in this case, does not provide any useful information. In the data, there is a lot of points, which makes it hard to properly presents them on the scatter plot. Another approach is to use kernel density estimate (kde) plots, also called density plots. Figure 9.4 presents the same UMAP embedding visualisation on kde plot.

But still a proper distinction between classes is not provided. While part of COVID images is slightly repulsed from the clump of clusters, clusters representing pneumonia

and normal images overlap almost entirely. This is also not coherent with medical expert knowledge. Images from normal class should be driven further away from both COVID and pneumonia. Clusters shown in Figure 9.4 result from UMAP based on probably inappropriate features.

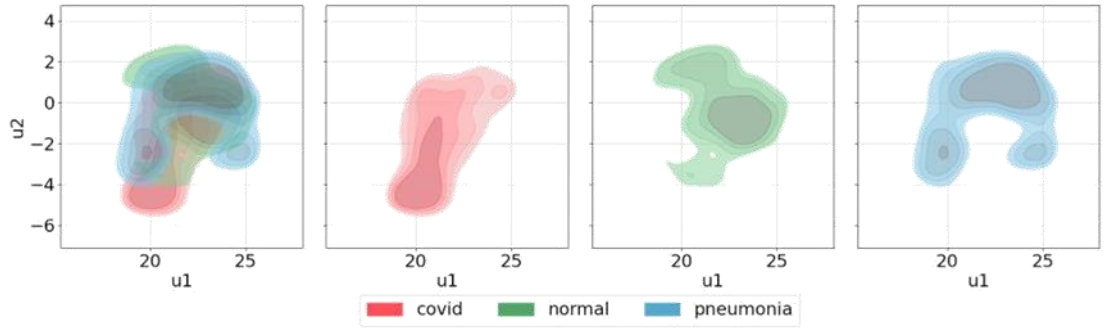

Fig. 9.4. Result of the UMAP procedure on the vectorized images
Rys. 9.4. Rezultat techniki UMAP wyliczonej na obrazach zwektoryzowanych – izolinie

9.3.2. UMAP on radiomics from whole images

To address issues occurring while calculating UMAP on vectorized images, we propose moving from the vectorized space into the feature space. Feature space is defined on the basis of radiomic features 2.2.

Initially, the whole image is considered as the ROI for radiomics calculations. Before the application of the UMAP procedure, features were pre-processed using pipeline described in subsection 2.3. The resulting embedding is presented in Figure 9.5.

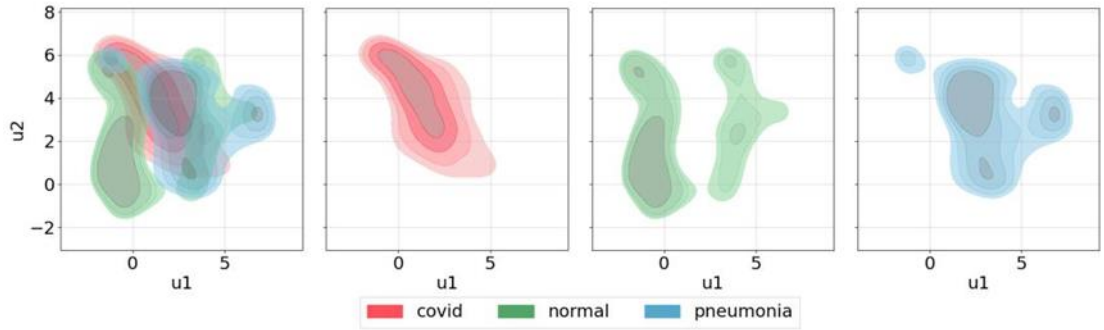

Fig. 9.5. Result of UMAP on radiomic features calculated from whole images
Rys. 9.5. Rezultat techniki UMAP wyliczonej na podstawie cech radiomicznych całych obrazów

As seen in Figure 9.5, there is a difference in the distribution of subsequent classes. While COVID images are forming one cluster, pneumonia and normal have multiple, scattered clusters.

9.3.3. UMAP on all radiomics from the lung regions

As the results are still not satisfying we propose to narrow down the ROI in order to gain access to the information more relevant to the problem.

According to [24–26] the most relevant information about the state of the patient are present within the lungs. Therefore, lung regions were extracted from the images and treated as a new ROI. The segmentation was done with the use of U-Net neural network architecture described in subsection 2.4. The results of the segmentation are shown in Figure 9.2b. The chosen radiomic features were calculated on the extracted lung regions and the UMAP procedure was conducted as before, with a change to the UMAP metric from ’euclidean’ to ’seuclidean’. Results are presented in Figure 9.6.

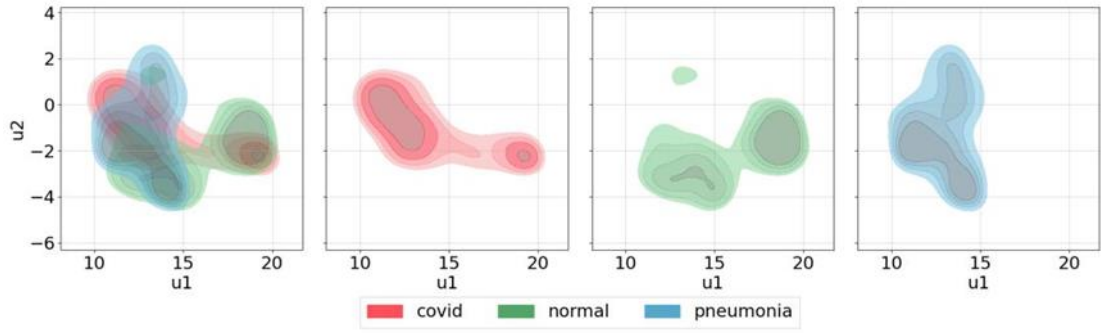

Fig. 9.6. Result of the UMAP procedure on lung radiomic features
Rys. 9.6. Rezultat techniki UMAP wyliczonej na podstawie cech radiomicznych wysegmentowanych płuc

The separation between classes is more visible but the number of clusters is overestimated. Also, a major part of normal class images overlaps with the smaller COVID cluster. Those results indicate that there still could be features that attract unrelated clusters towards each other.

9.3.4. UMAP on selected radiomics from the lung region

In contrast to [11–13] we are looking for information enveloped, hidden within the ROI and not directly indicated by it. In order to create UMAP embeddings more suited towards considered problem of normal, pneumonia and COVID distinction, we propose a semi-supervised embedding method with feature selection. The feature selection is based on effect size, 2.5 and described in subsection 2.6.

Three subsets of the features were considered based on the created ranking. Firstly, features with at least a small effect size ($\eta^2 > 0.01$) were taken into consideration. This

resulted in 71 radiomic features that were fed into UMAP. However, the results were not satisfactory, the number of features was too high. In the second attempt, a line connecting the highest and the lowest effect size values on the bar plot (Figure 9.7) was drawn. Euclidean distance from each bar to the line was calculated and two shortest distances were chosen and marked on the plot that pointed to the threshold cut-offs for the feature selection. With the first cut-off point ($\eta^2 > 0.04$), a set of 13 features was included and with the second cut-off point ($\eta^2 > 0.03$), a set of 30 features was included. For both feature sets the UMAP procedure was conducted.

While both, 13 and 30 features UMAP were created, the one with 30 features was selected. The parameters were same as in the previous section's UMAP.

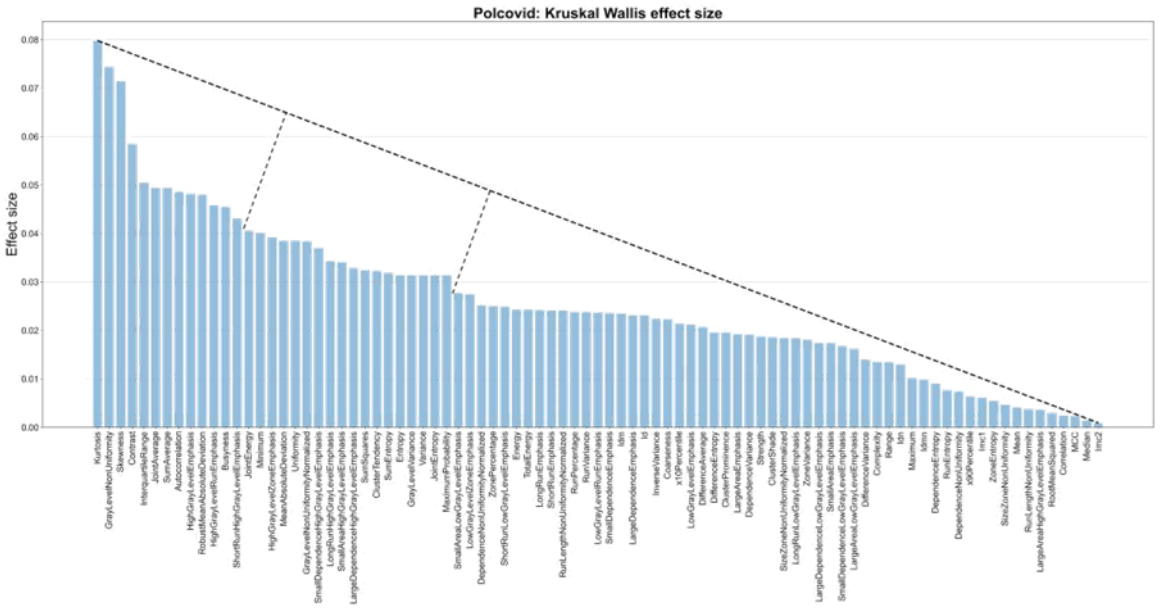

Fig. 9.7. Feature importance ranking based on calculated effect sizes
Rys. 9.7. Ranking ważności cech na podstawie obliczonych wielkości efektów

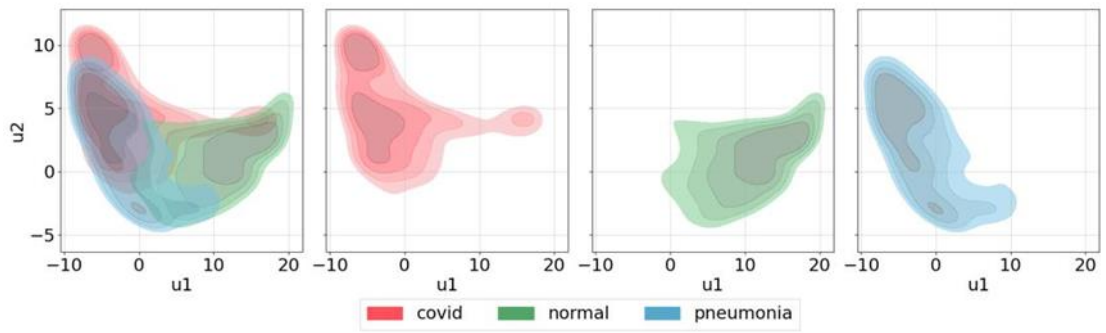

Fig. 9.8. Result of the UMAP procedure on lungs features
Rys. 9.8. Rezultat techniki UMAP wyliczonej na podstawie wybranych cech radiomicznych wysegmentowanych płuc

As shown in Figure 9.8, healthy (normal class) patients are clearly separated from the pneumonia and COVID patients. While COVID and pneumonia cluster overlaps, pneumonia has disjointed parts from the COVID cluster. COVID cluster concentration point lays partially beyond the pneumonia cluster and its edges have some independent parts.

9.4. Discussion

All figures shown up to this point, represented image embeddings created in order to differentiate disease entities. While the embeddings were shown in a form of density clusters, each cluster consisted of some points and each point indicated image form the POLCOVID data base. In order to compare proposed approach with a pure image vectorization, 12 example images from each class are shown on Figures 9.9 and 9.10. Images are taken from a vicinity of the most representative class density estimates.

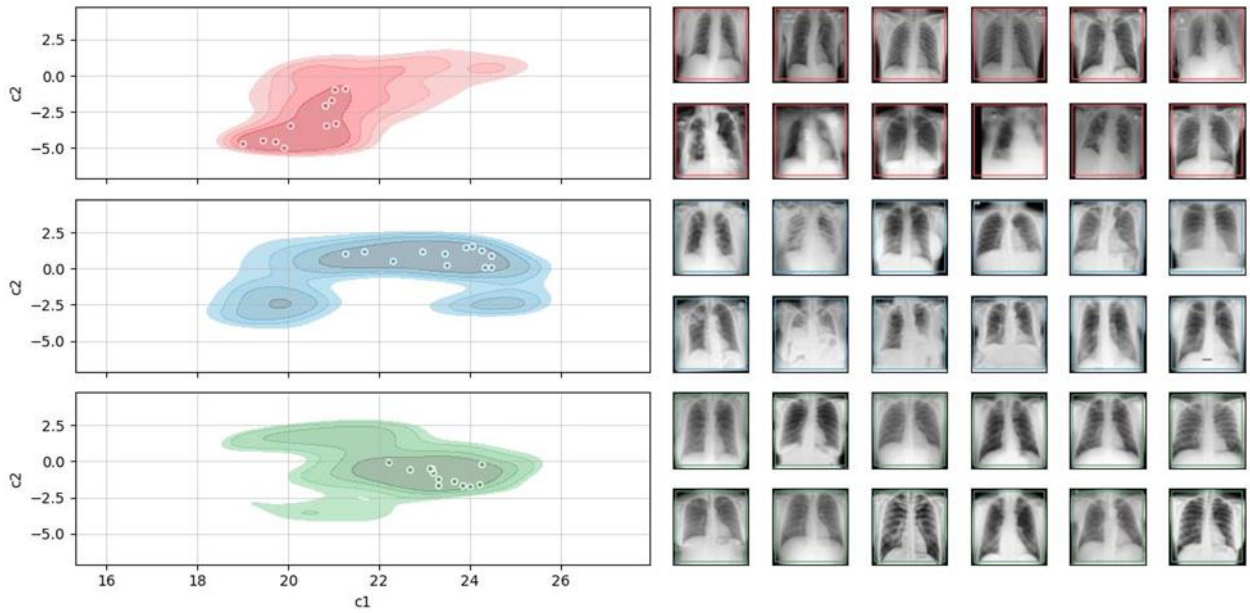

Fig. 9.9. Example images of each class from most prominent clusters of UMAP on vectorized images (4)
Rys. 9.9. Przykładowe obrazy dla każdej klasy z najbardziej widocznych klastrów UMAP na zwektoryzowanych obrazach

At a first glance it is difficult to infer on what basis the images are distributed. Due to its high cluster overlap (Figure 9.4) the prominence of the features characteristic for patient state is not obvious. According to UMAP original paper [1], a UMAP embedding is also meant to distribute clusters preserving their inter-cluster differences. Most representative COVID cluster seams to consist of some studies with relatively low pathological lung involvement mixed together with thous showing relatively high patholo-

gies. Cluster of pneumonia disease with quite significant lung changes seems to be to close to the cluster of normal studies. On the Figure 9.10 the distribution of the clusters is much easier to interpret. COVID cluster with prominent pathological lung involvement is driven further away from the normal cluster, which does not contains aggressive changes. Pneumonia cluster has its points distributed close to the COVID cluster indicating similarities in lung abnormality. Normal cluster distributes itself between two most prominent clump of points, both laying quite far from pneumonia and COVID clusters. All around, the final shape creates structure similar to letter ”U”. It seems to indicate a shift from the changes with most prominent pathological lung involvement (upper left corner,) to the one with the lowest pathological lung involvement (upper right corner).

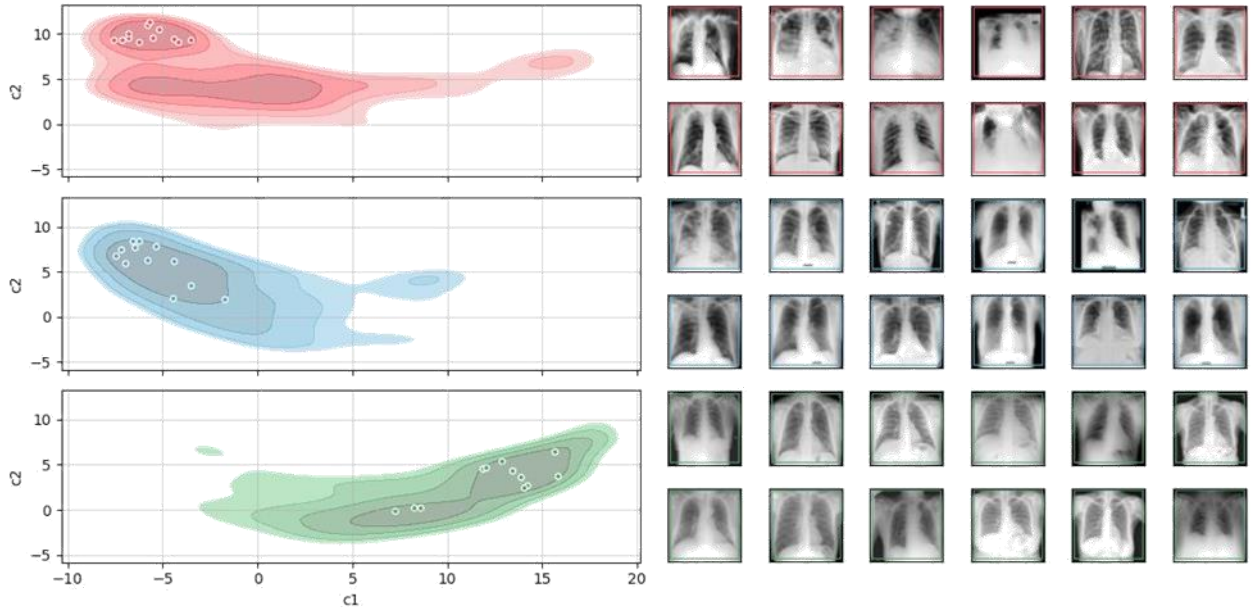

Fig. 9.10. Example images of each class from most prominent clusters of UMAP on lungs radiomic features (8)

Rys. 9.10. Przykładowe obrazy każdej klasy z najbardziej wyróżniających się skupień UMAP na wybranych cechach radiometrycznych płuc.

9.5. Conclusion

Obtained clusters in UMAP embedding are more readable and interpretative visualisation of disease distribution across the embedding, leading to much more meaningful conclusions.

POLCOVID study group

Department of Infectious Diseases and Hepatology, as coordinator: Jerzy Jaroszewicz (Medical University of Silesia in Katowice, Infectious Diseases Hospital No. 1 in Bytom), Jan Baron, Katarzyna Gruszczynska (Department of Nuclear Medicine and Image Diagnostics, Medical University of Silesia in Katowice), Magdalena Sliwinska, Mateusz Rataj, Przemyslaw Chmielarz (Voivodship Specialist Hospital in Wroclaw), Edyta Szurowska (II Department of Radiology, Medical University of Gdańsk), Jerzy Walecki, Samuel Mazur, Piotr Wasilewski (Central Clinical Hospital of the Ministry of Internal Affairs and Administration in Warsaw), Tadeusz Popiela, Justyna Kozub (Collegium Medicum of the Jagiellonian University in Kraków), Grzegorz Przybylski, Anna Kozanecka (Kujawsko-Pomorskie Pulmonology Center in Bydgoszcz), Andrzej Cieszanowski, Agnieszka Oronowicz-Jaskowiak, Bogumil Golebiewski (National Institute of Oncology in Warsaw, Department of Imaging Diagnostics), Complex of Health Care Centres, Mateusz Nowak (Silesian Hospital in Cieszyn), Barbara Gizycka (Single Infectious Diseases Hospital Megrez Ltd. in Tychy: Department of Imaging Diagnostics), Piotr Blewaska (District Hospital in Raciborz), Department of Infectious Diseases and Hepatology, University of M. Kopernika w Toruniu, Malgorzata Pawlowska, Piotr Rabiko, Pawel Rajewski (Collegium Medicum in Bydgoszcz), Department of Radiological and Imaging Diagnostics, Jerzy Walecki (Medical Center for Postgraduate Education, Warsaw), Clinical Department of Imaging Diagnostics, Katarzyna Sznajder (University Clinical Hospital in Opole), Department of Infectious Diseases University of Rzeszow, Robert Plesniak (Medical Center in Lancut), Department of Allergology and Internal Medicine, Marcin Moniuszko (Medical University of Bialystok), Department of Infectious Diseases and Hepatology, Robert Flisiak (Medical University of Bialystok), Andrzej Cieszanowski (Medical University of Warsaw: II Department of Clinical Radiology), Przemyslaw Bombinski (Department of Pediatric Radiology), Agata Majos (Medical University of Lodz: Department of Radiological and Isotopic Diagnostics and Therapy), Michal Mik (Department of General and Colorectal Surgery), Medical University of Wroclaw, Krzysztof Simon (Department of Infectious Diseases and Hepatology), Bartosz Markiewicz (Voivodship Comprehensive Hospital in Kielce: Department of Imaging Diagnostics), Gabriela Zapolska, Krzysztof Klaude, Katarzyna Rataj (Czerniakowski Hospital in Warsaw), Sebastian Hildebrandt, Katarzyna Krutul-Walenciej (Central Clinical Hospital of the Medical University of Gdansk), Adrianna Tur, Grzegorz Drabik (Prognostic Specialist Clinic in Knurow), Damian Piotrowski (Specialist Hospital No. 1 in Bytom).

Acknowledgement

The research leading to these results was partially funded from the National Science Centre, Poland, grant MNiSW/2/WFSN/2020 project name CIRCA – COVID-19 online image diagnostic support service. Calculations were carried out using GeCONiI infrastructure funded by NCBiR project no. POIG.02.03.01-24-099/13. Additionally, AS and WP are a holders of European Union scholarship through the European Social Fund, grant POWR.03.05.00-00-Z305.

Bibliography

1. Leland McInnes, John Healy, James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.
2. Ethem Alpaydin, Fevzi Alimoglu. Pen-based recognition of handwritten digits data set. 1996.
3. A. Sameer Nene, K. Shree Nayar, Hiroshi Murase. Columbia object image library (coil-20). Technical report, 1996.
4. Sameer A. Nene, Shree K. Nayar, Hiroshi Murase. Columbia object image library (coil-100). 1996.
5. Han Xiao, Kashif Rasul, Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.
6. Yingfan Wang, Haiyang Huang, Cynthia Rudin, Yaron Shaposhnik. Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization, 2020.
7. Abdelhakim Cheriet Mebarka Allaoui, Mohammed Lamine Kherfi. Considerably improving clustering algorithms using umap dimensionality reduction technique: A comparative study. In *Image and Signal Processing*, pages 317–325, Cham, 2020. Springer International Publishing.
8. Yann LeCun, Corinna Cortes. Mnist handwritten digit database. 2010.
9. Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.
10. Andrew C. Yang, Daniel J. Wu, Vinay U Prabhu. Afro-mnist: Synthetic generation of mnist-style datasets for low-resource languages, 2020.

11. Jeong Hoon Lee, Eun Ju Ha, Jin Roh, Su Jin Lee, Jeon Yeob Jang. Technical feasibility of radiomics signature analyses for improving detection of occult tonsillar cancer. *Scientific Reports*, 11(1):192, Jan 2021.
12. Mohammadhadi Khorrami, Kaustav Bera, Rajat Thawani, Prabhakar Rajiah, Amit Gupta, Pingfu Fu, Philip Linden, Nathan Pennell, Frank Jacono, Robert C. Gilkeson, Vamsidhar Velcheti, Anant Mad-abhushi. Distinguishing granulomas from adenocarcinomas by integrating stable and discriminating radiomic features on non-contrast computed tomography scans. 03 2021.
13. Andrea Bizzego, Nicole Bussola, Damiana Salvalai, Marco Chierici, Valerio Maggio, Giuseppe Jurman, and Cesare Furlanello. Integrating deep and radiomics features in cancer bioimaging. 03 2019.
14. Virendra Kumar, Yuhua Gu, Satrajit Basu, Anders Berglund, Steven A. Eschrich, Matthew B. Scha-bath, Kenneth Forster, Hugo J.W.L. Aerts, Andre Dekker, David Fenstermacher, Dmitry B. Goldgof, Lawrence O. Hall, Philippe Lambin, Yoganand Balagurunathan, Robert A. Gatenby, and Robert J. Gillies. Radiomics: the process and the challenges. *Magnetic Resonance Imaging*, 30(9):1234–1248, 2012. Quantitative Imaging in Cancer.
15. Robert J. Gillies, Paul E. Kinahan, Hedvig Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016. PMID: 26579733.
16. Vishwa Parekh, Michael A. Jacobs. Radiomics: a new application from established techniques. *Expert Review of Precision Medicine and Drug Development*, 1(2):207-226, 2016. PMID: 28042608.
17. Stephen Yip and Hugo Aerts. Applications and limitations of radiomics. *Physics in Medicine and Biology*, 61(13):R150–R166, jun 2016.
18. Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, Hugo J.W.L. Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107, 2017.
19. Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.
20. Olaf Ronneberger, Philipp Fischer, Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

21. Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks, 2017.
22. Jacob Cohen. Statistical power analysis for the behavioural sciences. hillsdale, nj: Laurence erlbaum associates, 1988.
23. *Kruskal-Wallis Test*, pages 288–290. Springer New York, New York, NY, 2008.
24. Wei Zhao, Zheng Zhong, Xingzhi Xie, Qizhi Yu, Jun Liu. Relation between chest ct findings and clinical conditions of coronavirus disease (covid-19) pneumonia: A multicenter study. *American Journal of Roentgenology*, 214(5):1072–1077, May 2020.
25. Adam Bernheim, Xueyan Mei, Mingqian Huang, Yang Yang, Zahi A. Fayad, Ning Zhang, Kaiyue Diao, Bin Lin, Xiqi Zhu, Kunwei Li, Shaolin Li, Hong Shan, Adam Jacobi, and Michael Chung. Chest ct findings in coronavirus disease-19 (covid-19): Relationship to duration of infection. *Radiology*, 295(3):200463, 2020. PMID: 32077789.
26. Shohei Inui, Akira Fujikawa, Motoyuki Jitsu, Naoaki Kunishima, Sadahiro Watanabe, Yuhi Suzuki, Satoshi Umeda, Yasuhide Uwabe. Chest ct findings in cases from the cruise ship diamond princess with coronavirus disease (covid-19). *Radiology: Cardiothoracic Imaging*, 2(2):e200110, 2020.

UMAP-BASED GRAPHIC REPRESENTATION OF POLCOVID CHEST X-RAY DATA SET HETEROGENEITY

Abstract

Visualisation is an essential step in the process of understanding the data. Characterisation of underlying structures is often difficult, especially when dealing with large number of dimensions. Uniform Manifold Approximation and Projection (UMAP) technique can solve this problem by transforming high-dimensional data to low-dimensional embedding. In image analysis, the dominant approach to define input of the UMAP procedure is to transform image into a single-column vector. This is justified for images where the key information is included in the dominant structures clearly visible on the image. In the case of chest X-ray imaging, in COVID-19 patients, the biological changes are subtle and hard to indicate in a single region of interest (ROI), but play a key role in medical diagnosis. We have analysed the data from POLCOVID database, which contains 4956 lung radiographs collected from 24 Polish hospitals. The dataset

includes radiographs of healthy subjects (n=2578), COVID-19 patients (n=1204) and pneumonia patients (n=1174). We propose a novel method of determining and selecting radiomic features to visualise significant differentiation of radiographs between patient groups in UMAP space. Our approach achieves significantly more interpretative UMAP embedding of the disease distribution.

Keywords: UMAP, chest X-Rays, COVID-19, radiomics, images.