DOI:10.34918/85106

Aleksandra SUWALSKA[1,*], Marek SOCHA[1], Wojciech PRAŻUCH[1],
Joanna TOBIASZ[1,2], Joanna POLAŃSKA[1], Michał MARCZYK[1,*]
and POLCOVID Study Group

Chapter 6. nUMAP: NEURAL NETWORK BASED UMAP SOLUTION FOR THE MULTI DATASET VISUALISATION

6.1. Introduction

High-dimensional data is common in the field of biomedicine. It is difficult to analyze, therefore it is hard to find its flaws and hidden relations. Dimensionality reduction techniques help researchers to overcome this problem by embedding high dimensional information into the lower-dimensional space. Such embeddings can be visualised and then analysed properly.

One of such embedding techniques is Uniform Manifold Approximation and Projection (UMAP) [1]. Besides embedding it can also be used for clustering and data pre-processing. This method thrives when used with tabular data as an input and despite that it can be used with image data [3-5], it has limited use cases. The common factor of referenced papers is the usage of homogenous image datasets with neutral backgrounds. It is very rare to stumble upon such data in biomedicine.

In the paper [2], a solution to the problem of image data heterogeneity was proposed by introducing a novel method capable of robust transformation of X-ray radiogram into a set of features and UMAP based pipeline capable of embedding features into low-dimensional space. While the method is valid for X-ray images, it requires a definition of a region of interest (ROI). For other biomedical datasets, this may not be possible. In the study, some modifications are introduced to unify the previously

[1] Department of Data Science and Engineering, Faculty of Automatic Control, Silesian University of Technology, Gliwice, Poland.
[2] Department of Graphics, Computer Vision and Digital Systems, Faculty of Automatic Control, Silesian University of Technology, Gliwice, Poland.
* Corresponding authors: aleksandra.suwalska@polsl.pl, michal.marczyk@polsl.pl.

proposed method. The proposed pipeline consists of neural network (NN) guided features extraction and UMAP embedding followed by NN universal embedding learning.

Guided feature extraction relies on the information obtained from the latent vector of a pre-last layer of a neural network train specifically for this task [6] (latent space with UMAP in the context of fashion recommendation system), [7] (using latent space vector for UMAP in clustering of genes). UMAP procedure embeds latent space features into the two-dimensional space allowing visualisation and relation analysis. To achieve robust features dependence a regression neural network was trained which learnt the embedding. The pipeline results in a method capable of dealing with numerical, image and mixed data types. The goal of the study was to present different use cases of the proposed method, like discovering the batch effect, analysing the dataset's quality, and explaining the neural network prediction, in order to prove its wide applicability.

6.2. Datasets

6.2.1. Mass cytometry dataset

In the study, a mass cytometry dataset was used that comprised two healthy control samples with the number of bronchoalveolar lavage cells (BALC) equal to 329,228 and 341,007, respectively. The samples come from studies on drug-resistant tuberculosis. Bronchoscopies were performed in the bronchoscopy theatre, ward A5, Tygerberg Hospital (TBH) in Cape Town, South Africa. The cells' signal was measured with the CyTOF2 instrument, at the South African Tuberculosis Vaccine Initiative at the University of Cape Town. For each cell, a set of 32 markers was collected. The dataset was preprocessed (pre-gated) and arcsinh transformed with a co-factor of 5.

6.2.2. Chest X-Ray dataset

The chest X-Ray image dataset was composed of POLCOVID database [2] and COVIDx database [8]. The data from POLCOVID database were collected from 24 Polish hospitals (see POLCOVID Study Group section) during a CIRCA project and

consists of 2426 healthy patients, 1147 patients with pneumonia and 1236 patients with COVID-19 (with positive RT-PCR test results). The COVIDx database consists of 8066 healthy patients, 5573 patients with pneumonia and 1763 patients with COVID-19. The images were resized to 512x512 pixels resolution, had their lungs segmented from an images and were scaled.

6.3. nUMAP

The proposed nUMAP method is a modification of the standard UMAP approach (Fig. 1). Since UMAP only accepts numerical data as input, the limitation can be solved with a neural network that can take different types of data as input. The final fully-connected layer provides feature vectors that are combinations of the input data in a numerical form that can be further analysed by the standard UMAP model.

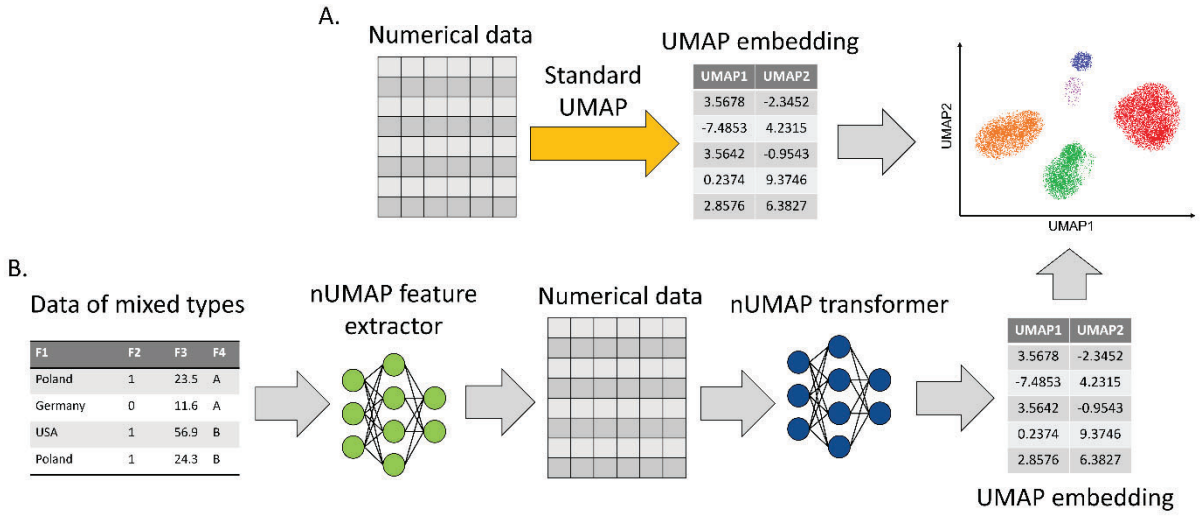

Fig. 1. nUMAP pipeline. A) Standard UMAP model trained on numerical data to generate its embedding that can be visualized in 2D space. B) nUMAP accepts data of mixed types that are processed into numerical representation by the first neural network (nUMAP feature extractor). Then the second neural network transforms the numerical data into embedding, giving identical results to the standard UMAP approach

Rys. 1. Schemat działania nUMAP: A) Standardowa metoda UMAP trenowany na danych numerycznych w celu zwizualizowania ich zredukowanej reprezentacji w przestrzeni 2D. B) nUMAP akceptuje dane o mieszanych typach dzięki zastosowaniu sieci neuronowej (ekstraktor cech nUMAP), która przetwarza je na dane numeryczne. Następnie druga sieć neuronowa transformuje dane numeryczne do reprezentacji o niższej wymiarowości, która jest identyczna do rezultatu otrzymanego ze standardowej metody UMAP

The first neural network is called the nUMAP feature extractor which enables simultaneous processing of different types of data. The nUMAP feature extractor can be any model that can process desired input data into numerical features that can be further analysed. For example, like in the study, it can be a convolutional neural network that accepts images and numerical features as inputs to combine them into new numerical features.

The second part of the nUMAP approach is another neural network, the nUMAP transformer, that learns how to map numerical features into UMAP embeddings. The network is a multilayer perceptron for regression problems, that accepts numerical data as input and UMAP embedding as output. The transformer's architecture used in the study was a simple neural network with two hidden layers consisting of 100 and 50 neurons, respectively. The number of input neurons was equal to the number of numeric features and the number of the output neurons was two (for the reduction into 2D space embedding). The learning rate was set to 0.001.

6.4. nUMAP application examples

6.4.1. Use case 1: visualization of a batch effect in mass cytometry data

Mass cytometry datasets are numerical, therefore the standard UMAP approach can be applied as presented in Fig. 2. The embedding is created and can be visualized in 2D space. The pink area represents the region with the highest density of cells (50% of data) from the sample in reference to all cells (grey points).

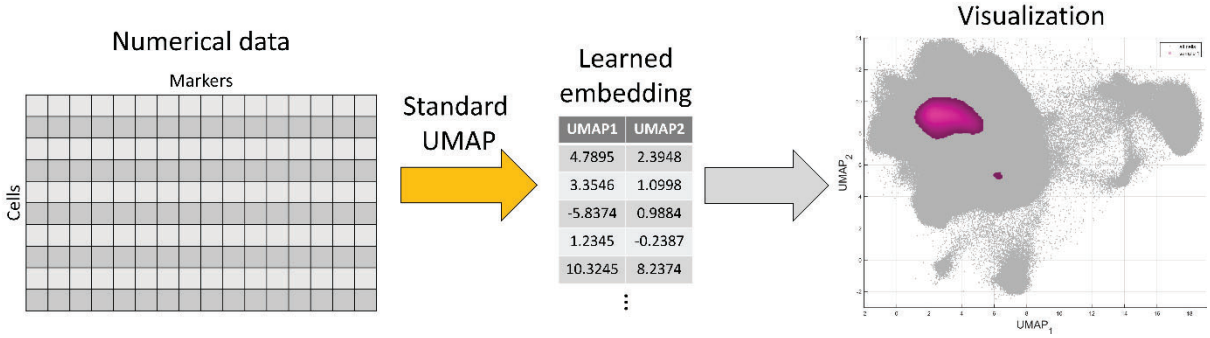

Fig. 2. Standard UMAP approach for mass cytometry data. One sample is transformed into embedding that allows visualizing the dataset in 2D space. Visualization: grey points – all cells; pink area – region with the highest density of cells

Rys. 2. Standardowa metoda UMAP dla danych z cytometrii masowej. Jedna próbka jest przetworzona do dwuwymiarowej reprezentacji, którą można zwizualizować na wykresie. Wizualizacja: szare punkty – wszystkie komórki; różowy region – obszar z największym zagęszczeniem komórek próbki

The problem appears when adding another sample from the same experiment to the first one to represent new data in the learned UMAP embedding. The existing UMAP method that transforms new data with the learned model is very time-consuming for big data. Since mass cytometry datasets may have tens of millions of records (cells), the standard UMAP transformer is inefficient. Therefore, the second part of the nUMAP system applies here as presented in Fig. 3. The trained nUMAP transformer generates embedding quickly and accurately.

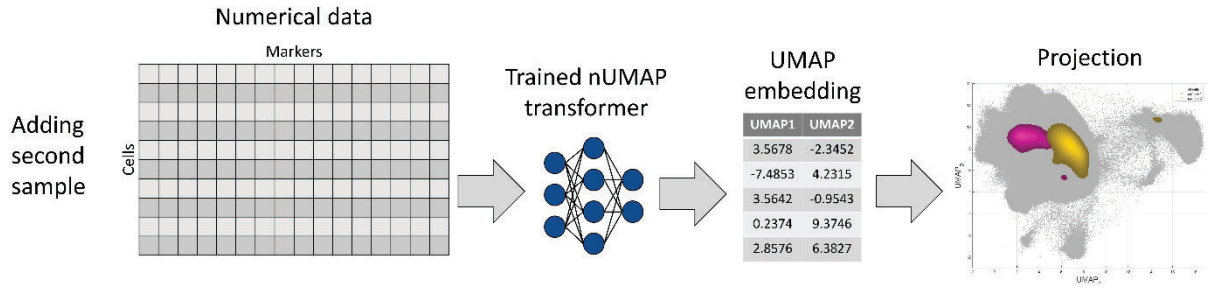

Fig. 3. nUMAP transformer allows fast embedding creation and projection for mass cytometry (big data). Visualization: grey points – all cells; pink area – a region with the highest density of cells from the first sample; yellow area – a region with the highest density of cells from the second sample that was transformed into the same UMAP space as the first sample

Rys. 3. Transformer nUMAP pozwala na szybkie wygenerowanie dwuwymiarowej reprezentacji danych z cytometrii masowej (duże dane). Wizualizacja: szare punkty – wszystkie komórki; różowy region – obszar z największym zagęszczeniem komórek próbki pierwszej; żółty region – obszar z największym zagęszczeniem komórek próbki drugiej transformowanej do tej samej przestrzeni UMAP co próbka pierwsza

Adding the second sample reveals the presence of a batch effect in the dataset. The batch effect is a technical variation in the data that makes it difficult to reveal biological relationships and should be removed or decreased. Many solutions have been proposed for batch effect removal for mass cytometry. nUMAP makes it possible to visualize the effect of such a batch correction algorithm (Fig. 4).

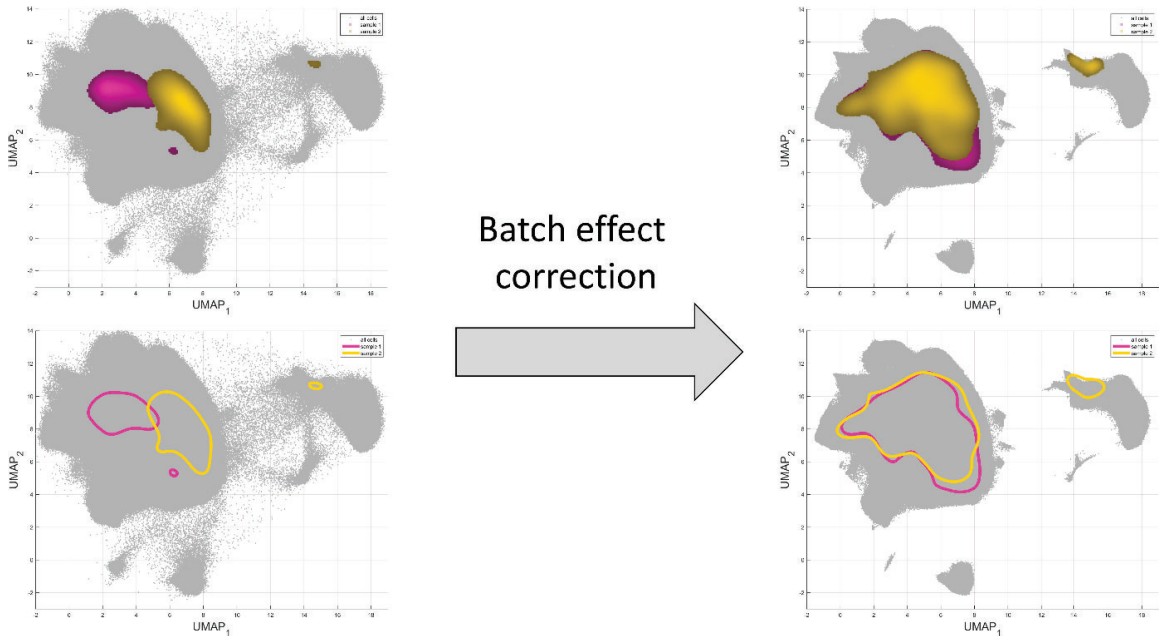

Fig. 4. Visualization of the mass cytometry dataset comprised of two samples before and after batch effect correction (pink and yellow areas representing the regions of their highest cell concentration – for better visualization). After batch effect correction a new UMAP embedding and nUMAP transformer are learned based on the corrected feature values. Since the samples overlap after the batch correction, it is visible that the batch effect was significantly reduced

Rys. 4. Wizualizacja danych z cytometrii masowej złożonych z dwóch próbek, przed i po korekcie efektu paczki (różowe i żółte obszary reprezentują regiony o największym zagęszczeniu komórek – dla lepszej wizualizacji). Wykorzystując wartości po korekcji efektu paczki wytrenowano nowy model nUMAP do stworzenia nowej reprezentacji UMAP. Ponieważ po korekcie próbki nakładają się na siebie, można wywnioskować że efekt paczki został znacząco zredukowany

6.4.2. Use case 2: RTG quality assessment

The X-Ray chest dataset consists of image data (radiograms) and clinical data. The goal was to validate whether the dataset contains hidden biases and flaws which are not related to the lung disease. Therefore a neural network was created to classify the data into three classes: normal (healthy), pneumonia and COVID-19. To verify data robustness, the nUMAP was employed as in Fig. 5. The neural network was used as the nUMAP feature extractor.

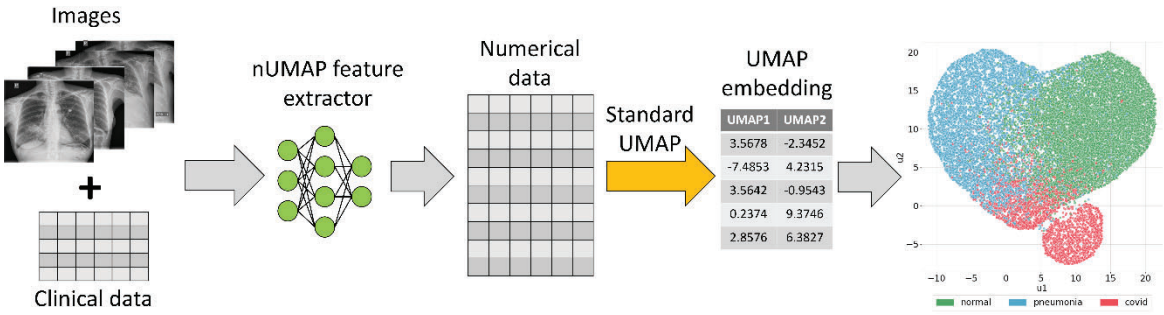

Fig. 5. Image information is merged with clinical data through the nUMAP feature extractor, and the standard UMAP method (with cosine distance metric) is used to project received numerical features on the 2D plane

Rys. 5. Informacja obrazowa jest połączona z danymi klinicznymi z wykorzystaniem ekstraktora cech nUMAP, następnie standardowa metoda UMAP (z metryką odległości cosine) transformuje otrzymane cechy numeryczne do przestrzeni o niższej wymiarowości

Ideally, UMAP embedding would be of compact shape with three fuzzy borders splitting the data into the three classes. While most of the data points follow mentioned behaviour, there is a group of COVID-19 data points which are visibly distinct from the other representatives of the category, forming a separate aggregate of points. Further analysis revealed that data within this 'island' consist of radiograms with relatively low original resolution (Fig. 6).

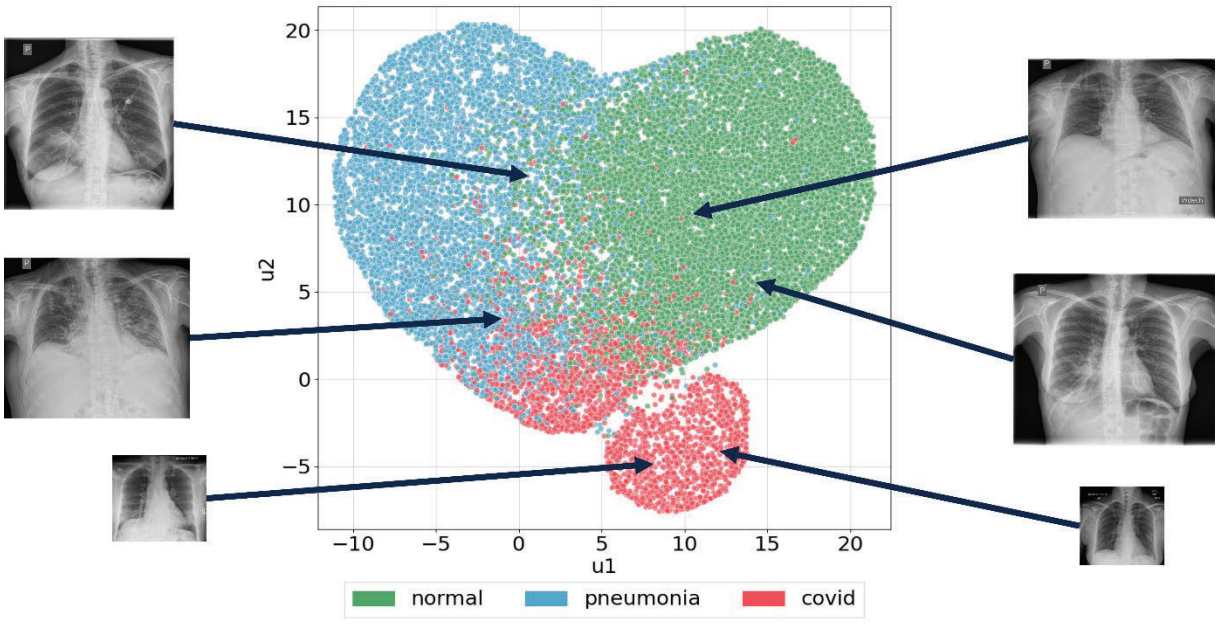

Fig. 6. Example radiograms representing embedded data points on the UMAP 2D plot. The radiograms were scaled following their original proportions to each other. The two smallest images of COVID-19 patients come from a separate aggregate of points

Rys. 6. Przykłady obrazów reprezentujących wartości UMAP po redukcji wymiarowości na wykresie dwuwymiarowym. Radiogramy zostały wyskalowane względem siebie zachowując oryginalne proporcje. Dwa najmniejsze obrazy pacjentów z COVID-19 pochodzą z osobnego skupiska punktów

A resolution of a radiogram has a big impact on its quality. Lower radiogram resolution results in a lower number of pixels which represents lung abnormal changes. Since the neural network input image resolution was 512 x 512 pixels, it was assumed that the island of points is mainly composed of images below the 512 x 512 resolution. To verify the thesis, points on the UMAP visualization were coloured according to the calculated radiogram resolutions (Fig. 7).

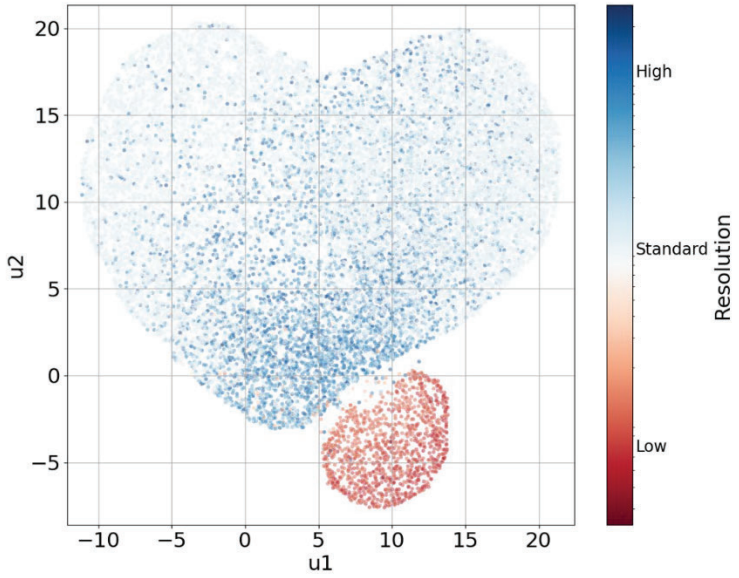

Fig. 7. UMAP embedding representation of X-ray dataset with colour scale pointing to radiogram resolution ranging from about 100x100 to nearly 4000x4000 pixels. Shown on a logarithmic scale

Rys. 7. Reprezentacja przestrzeni UMAP zdjęć rentgenowskich płuc z kolorową skalą wskazującą na rozdzielczość obrazów od około 100x100 do prawie 4000x4000 pikseli. Skala logarytmiczna

The radiograms contained within the smaller aggregate of points are of much lower resolution than radiograms in the compact part of UMAP embedding. This indicates that the neural network converged on the resolution rather than the disease entity. Showing that the problem is hidden in the dataset and calls for data curation.

6.4.3. Use case 3: Explainability of a classifier's prediction

nUMAP can work like an Explainability AI method to clarify the model's prediction. This behaviour is used in the CIRCA classification system [9]. CIRCA nUMAP allows for the projection of new data points, that are analysed through the CIRCA portal, on the learned UMAP visualization. The learned UMAP was created with the use of X-Ray chest images consisting of three patient categories: normal, pneumonia and COVID-19. The nUMAP training process is presented in Fig. 8.

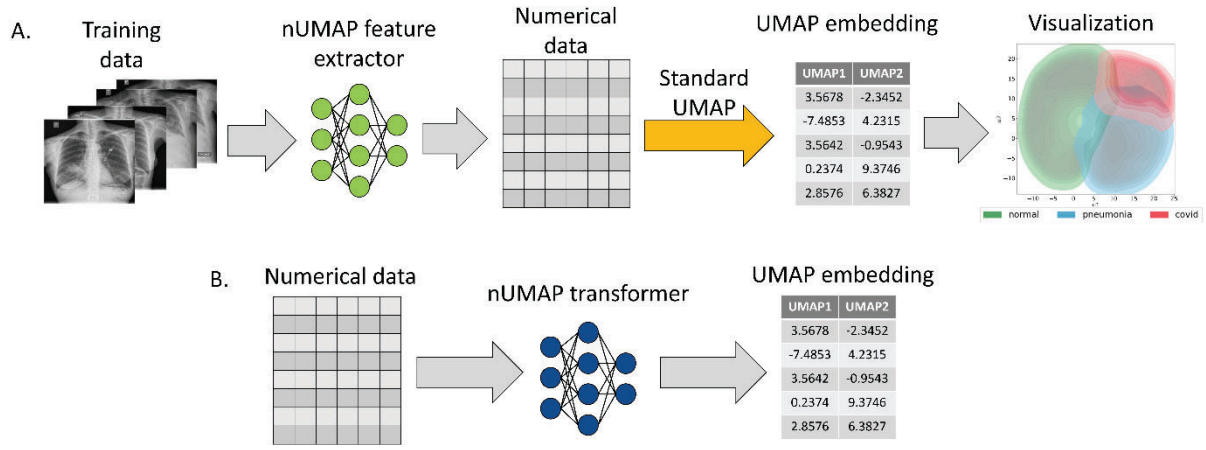

Fig. 8. Training of the CIRCA nUMAP system. A) Training dataset of X-Ray chest images is used to train the feature extractor. Then numerical features are transformed with the standard UMAP into embedding to create a 2D visualization. B) Numerical data extracted from the nUMAP feature extractor are used to train the nUMAP transformer to generate the same UMAP embedding as the standardapproach

Rys. 8. Trening metody nUMAP systemu CIRCA. A) Zbiór treningowy złożony z obrazów RTG klatki piersiowej jest wykorzystany do treningu ekstraktora cech nUMAP. Następnie cechy numeryczne są poddane transformacji standardową metodą UMAP do dwuwymiarowej reprezentacji. B) Dane numeryczne otrzymane z nUMAP ekstraktora cech są wykorzystane do treningu nUMAP transformatora w celu wygenerowania takiej samej reprezentacji UMAP jak z metody standardowej

When a new chest X-Ray image is loaded into the CIRCA portal, it undergoes a series of preprocessing and classification steps. The goal is to classify the image into the three mentioned categories based on the visual markers. Since the task is difficult due to the heterogeneous nature of COVID-19 changes, displaying the resulting category may not be sufficient. nUMAP tries to explain the prediction by projecting the image into the UMAP visualization that is comprised of three different areas indicating the categories. Moreover, nUMAP revealed subcategories within each category that differentiate the patients based on their degree of advancement of pulmonary changes. The projection process is visualized in Fig. 9.

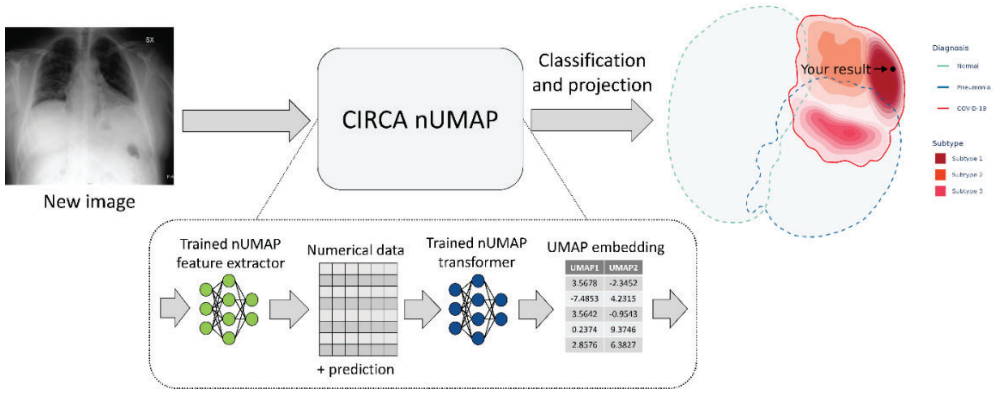

Fig. 9. A new image is processed by CIRCA system that consists of an nUMAP part responsible for projecting the analysis result into learned UMAP representation to visualize it in relation to the training samples. The image is classified as COVID-19 therefore the embedding is placed in the red area (regions indicating normal and pneumonia patients are shaded for better visualization). It can be seen that the result is placed in Subtype 1 of the COVID-19 category, which is the farthest from normal and pneumonia areas – therefore this is the subtype that contains patients with typical COVID-19 changes in the lungs

Rys. 9. Nowe zdjęcie RTG jest przetwarzane przez system CIRCA, który zawiera część nUMAP odpowiedzialną za projekcję do reprezentacji UMAP wyniku analizy w celu jego wizualizacji względem zdjęć ze zbioru treningowego. Obraz został zaklasyfikowany do kategorii COVID-19, dlatego jego dwuwymiarowa reprezentacja została umieszczona w rejonie czerwonym (regiony dla kategorii zdrowe płuca oraz zapalenie płuc zostały zacienione dla lepszej wizualizacji). Wynik został umieszczony w rejonie Podtypu nr. 1 kategorii COVID-19, który leży najdalej od obszarów kategorii zdrowe płuca i zapalenie płuc – w związku z tym jest to podtyp który zawiera obrazy pacjentów z typowymi zmianami COVID-19 w płucach

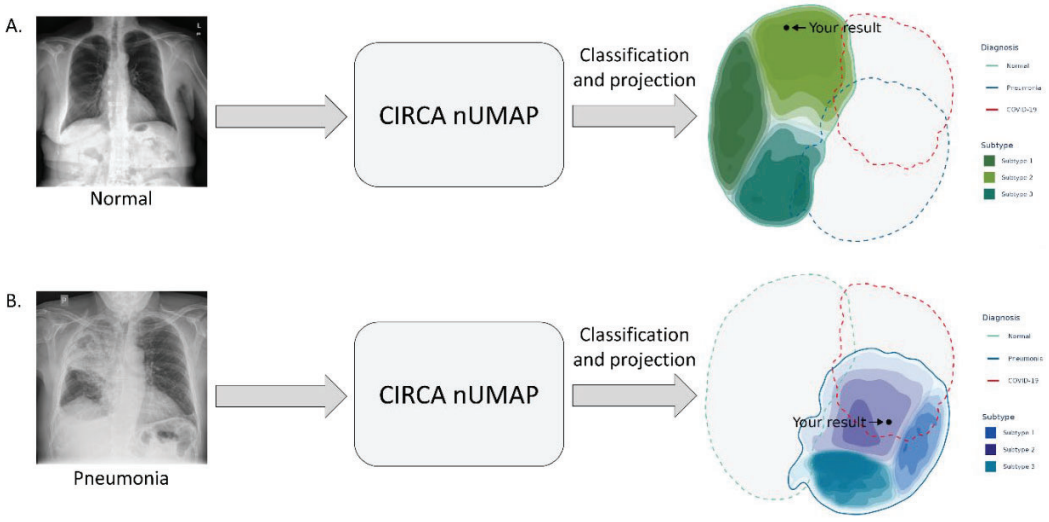

Fig. 10. Examples from CIRCA portal projection of the RTG analysis result in UMAP representation. A) Chest X-Ray image of normal (healthy) lungs is classified and projected with nUMAP into the healthy region of the embedding visualization. B) Chest X-Ray image of lungs with pneumonia changes is classified and projected with nUMAP into the pneumonia region of the embedding visualization

Rys. 10. Inny przykład z systemu CIRCA pokazuje projekcje Przykłady z portalu CIRCA, projekcja wyniku analizy RTG w reprezentacji UMAP. A) Obraz rentgenowski klatki piersiowej prawidłowych (zdrowych) płuc jest klasyfikowany i rzutowany za pomocą nUMAP do rejonu osadzenia kategorii zdrowych. B) Obraz rentgenowski płuc ze zmianami zapalenia płuc jest klasyfikowany i rzutowany za pomocą nUMAP do rejonu osadzenia kategorii zapalenia płuc

6.5. Discussion

Big data analysis is a demanding task that requires appropriate and efficient techniques to gain meaningful insights and results. This often requires presenting data in a lower-dimensional feature space. Methods like UMAP allow for dimension reduction preserving hidden structures in the data, however, they have some limitations.

In the study, a modification of the standard UMAP approach was proposed, that overcomes its limitations. The nUMAP parts can be linked with the standard UMAP approach or customized to a specific problem, as needed.

Firstly, the proposed nUMAP transformer can be trained on numerical features and embedding values from the standard UMAP approach to generate identical embedding but more effectively, especially in the case of big data, like mass cytometry. It solves the problem of the standard UMAP projection of new points into the existing low--dimensional representation which is time-consuming.

Secondly, nUMAP accepts mixed types of data as input due to the use of a neural network that works as a feature extractor. Therefore, different data types can be combined together into a numerical representation that can be used either with the standard UMAP approach or the nUMAP transformer to generate data embeddings. The shown examples of X-Ray data projection prove the usefulness of the approach.

The presented use cases of the nUMAP allowed for deeper insight into the data, revealing problems like batch effect in mass cytometry data or low-quality images of X-Ray dataset that have a great impact on further analysis results. Moreover, the method can be used to explain the classifier's prediction and therefore the usefulness of the trained classifier.

However, the nUMAP approach also has limitations. Since the feature extractor is a classification network, the method is no longer fully unsupervised and requires some knowledge about the labels (categories) of observations. Future work could focus on the use of a different feature extraction method that accepts mixed types of data as inputs.

6.6. Conclusion

nUMAP is a fast and effective method for embedding generation that can help visualize modifications of big data of mixed types. nUMAP overcomes the standard UMAP limitations and can be used for various purposes.

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 by the National Science Centre, Poland, grant MNiSW/2/WFSN/2020 project name CIRCA – COVID-19 online image diagnostic support service. MM was financed by grant no. 02/070/BK_22/0033. WP and JP were financed by OPUS grant no. 2017/27/B/NZ7/01833. Additionally, AS and WP are holders of a European Union scholarship through the European Social Fund, grant POWR.03.05.00-00-Z305 and JT is a holder of scholarship grant POWR.03.02.00-00-I029. Calculations were carried out using GeCONiI infrastructure funded by NCBiR project no. POIG.02.03.01-24-099/13).

Bibliography

1. L. Mclnnes, J. Healy, J. Melville: Umap: Unifold manifold approximation and projection for dimension reduction, *arxiv preprint* (2018).
2. M. Socha, A. Suwalska, W. Prazuch, M. Marczyk, J. Polanska: UMAP-based graphic representation of POLCOVID chest X-Ray data set heterogeneity, *Recent Advances in Computational Oncology and Personalised Medicine* (2021) **1**:100–114.
3. H. Xiao, K. Rasul, R. Vollgraf: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, *arxiv preprint* (2017).
4. Y. Wang, H. Huang, C. Rudin, Y. Shaposhnik: Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, Atrimap, and pacmap for data visualization, *arxiv preprint* (2020).
5. M. Allaoui, M.L. Kherfi, A. Cheriet: Considerably improving clustering algorithms using UMAP dimensionality reduction technique: a comparative study. *In International Conference on Image and Signal Processing, Springer, Cham* (2020) 317–325.

6. A.R. Sulthana, M. Gupta, S. Subramanian, S. Mirza: Improvising the performance of image-based recommendation system using convolution neural networks and deep learning. *Soft Computing* (2020) **24(19)**:14531–14544.
7. J. Ding, A. Regev: Deep generative model embedding of single-cell RNA-Seq profiles on hyperspheres and hyperbolic spaces, *Nature Communications* (2021) **12**:2554.
8. L. Wang, Z.Q. Lin, A. Wong: COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci Rep* (2020) **10**:19549.
9. WWW: https://circa.aei.polsl.pl/, access: 16.05.2022.

nUMAP: NEURAL NETWORK BASED UMAP SOLUTION FOR THE MULTI DATASET VISUALISATION

Abstract

High-dimensional data is associated with complex analysis and interpretation of the results. The problem is particularly relevant to biomedical problems like the analysis of cell expression profiles or medical imaging data. Big data contains potential noise that can cause the lack of optimal solutions. One way to deal with the problem is dimensionality reduction and feature selection. However, most of the methods are designed to process numerical data and for images, the methods require vectorization that amplifies the artefacts. Moreover, the techniques accept only one type of data at a time which could be insufficient to discover the real relationships. In the study, a novel method is proposed that is based on the UMAP dimension reduction technique. nUMAP combines the UMAP transformation with neural networks (NN) allowing for the processing of big data of mixed types. The method is based on a sequence of NN-UMAP-NN operations that extract features and create an embedding for effective visualization and inspection of the results. Moreover, the method can be applied to new data without the need for retraining. In the study, three real-world use cases of the nUMAP are presented: detection and correction of a batch effect in mass cytometry data, an inspection of the quality of chest X-Ray images and an explanation of a classifier's prediction. The work proves the effectiveness and wide application of the nUMAP.

Keywords: visualization, UMAP, neural network, embedding, big data