Data Visualisation by UMAP

Faouzi Alaya Cheikh, Mohib Ullah, Antoine Duperron

09/01/2023

Contents

1 Abstract :

1.1 Data :

The main objective is to discern distribution patterns in data from patients with three distinct diseases. This research is an integral part of a wider European initiative aimed at preventing the risk of falls among vulnerable people.

The sensor is placed in the patient's sole, and positioned to capture a range of measurements as the patient moves. These sensors enable continuous, non-intrusive monitoring, which is particularly important for sensitive populations prone to falls. The set of data collected encompasses a spectrum of measurements that reflect the physiological characteristics of patients suffering from their respective diseases (Parkinson's Disease, Stroke, Multiple scelorsis). Here we have 3 diseases, which can be interpreted as 3 different classes.

The final goal is to visualize the separation of the data between each class in order to know if it is possible to classify them well with the use of deep learning.

1.2 UMAP :

In the realm of data analysis, the emergence of high-dimensional data necessitates sophisticated techniques for comprehension. Uniform Manifold Approximation and Projection (UMAP) has garnered attention as a potent dimensionality reduction algorithm. UMAP uniquely preserves both local and global structures in diverse fields, from computational biology to image analysis.

UMAP assumes high-dimensional data lie on a lower-dimensional manifold. Constructing a weighted graph based on fuzzy topological relationships, UMAP maintains both local and global structures during embedding. Unlike t-SNE, UMAP balances global and local preservation through repulsive and attractive forces, revealing intricate patterns for exploratory analysis.

Applied in bioinformatics for single-cell RNA sequencing, UMAP identifies hidden cell subpopulations. In image analysis, UMAP enhances tasks like denoising and style transfer. With coherent visualizations, it proves pivotal for data interpretation.

This paper delves into UMAP's mathematical foundations, algorithmic workflow, strengths over existing methods, and its application in complex datasets. Providing a comprehensive understanding, this work empowers researchers and practitioners to leverage UMAP for enriched data analysis.

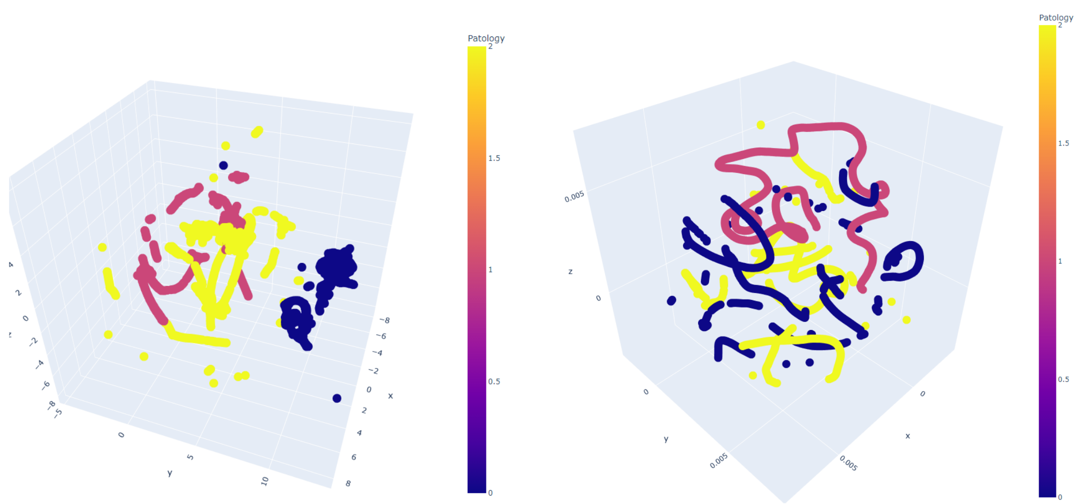
2 UMAP Application :

2.1 Setting of UMAP parameters :

First of all, you should know that UMAP can be used in two different ways, one supervised (we do a learning process, to create a projector that will allow better grouping of similar data during inference). And, we can also use this algorithm without learning (unsupervised).

Their's some parameters to fix before that :
- n-components : number of dimension of the result,
- n-neighbors : number of neighbors use,
- min-dist : distance available between the new data and his closer neighbor,
- random-state = 42 : Eliminates stochastic calculations,
- metric = 'euclidean' : we use euclidean distance during the process
- spread : coefficient which makes it possible to manage the propagation during the epochs, of the data in the new dimension
- n-epochs : number of epochs

2.2 Which use of UMAP ?



(a) Unsupervised UMAP result (b) Supervised UMAP result

Figure 1: Comparison between unsupervised / supervised umap application

We see here with this comparison that the unsupervised use of UMAP allows better clustering of classes so for the rest of this publication we will remain on unsupervised use.

2.3 Visualisation of diferents diseases :

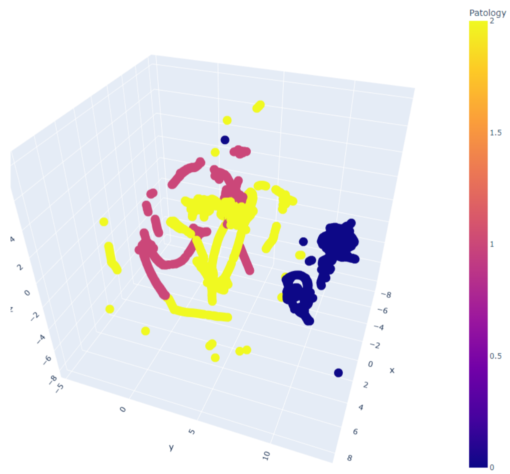


Figure 2: 3D result

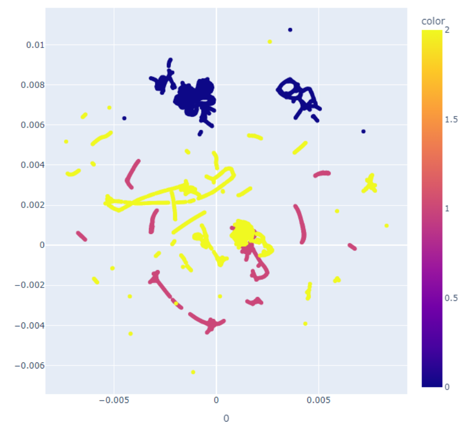


Figure 3: 2D result

Blue = MS (Multiple Sclerosis)
Pink = PD (Parkinson's Disease)
Yellow = Stroke

We see that the data of the MS class are well grouped compared to the PD and Stroke class which intertwine

2.4 Visualisation of Multiple Sclerosis patients :

We also aim to cluster the data of each of the patients who are suffering from the same disease, as if we had subclasses which would be the patient.

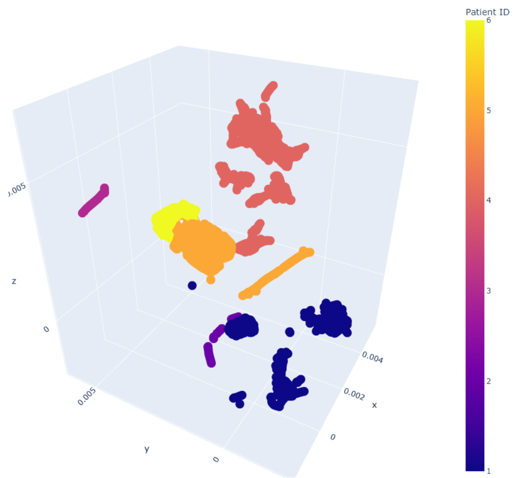


Figure 4: 3D result

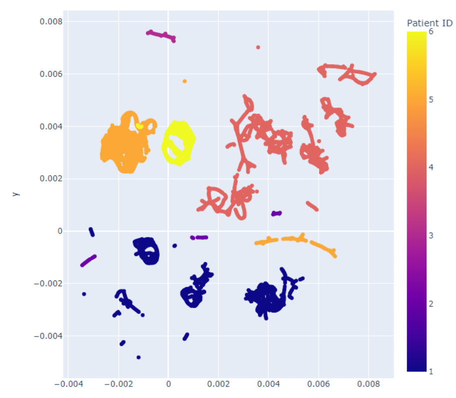


Figure 5: 2D result

We see for the patient n°2 (Purple) the data are too spread aroud the map, this is a problem for the clustering.

2.5 Visualisation of Parkinson's Disease patients :

We also aim to cluster the data of each of the patients who are suffering from the same disease, as if we had subclasses which would be the patient.

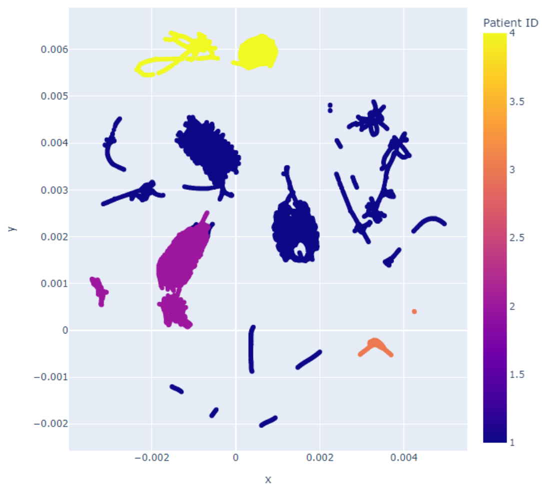


Figure 6: 3D result

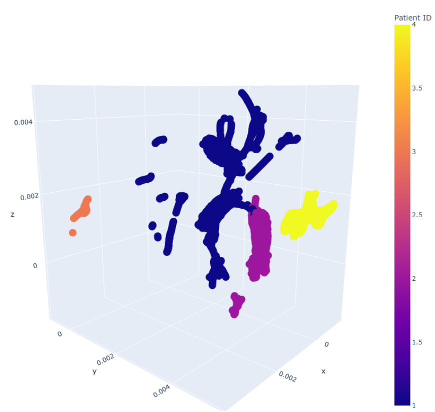


Figure 7: 2D result

We got a intertwine between the sub-class purple and blue one. And the blue sub-class is too spread in map.

3 Conclusion

In this study, we applied the UMAP algorithm to visualize and analyze high-dimensional data obtained from ankle-level sensors. The dataset encompassed measurements from patients with three distinct diseases: Multiple Sclerosis (MS), Parkinson's Disease (PD), and Stroke. The goal was to observe how UMAP's dimensionality reduction could reveal patterns and separations between these classes.

Our results demonstrated the effectiveness of UMAP in uncovering meaningful structures within the data. Through 2D and 3D visualizations, we observed clear groupings for the MS class, while the PD and Stroke classes exhibited more overlap. This distinction suggests that UMAP can aid in identifying data clusters corresponding to different diseases.

Furthermore, the unsupervised application of UMAP outperformed the supervised counterpart, emphasizing its potential in exploratory analysis. While this study has provided valuable insights, further research could focus on refining the clustering of subclasses, such as the spread of data within the maps.

The application of UMAP to medical data, as presented in this paper, opens avenues for better understanding disease patterns and facilitating early diagnostics. By leveraging the capabilities of UMAP, we contribute to the broader efforts of fall risk prevention in vulnerable populations, fostering advancements in healthcare and enhancing patient well-being.

Acknowledgments

We would like to express our gratitude to the European research initiative for providing the dataset used in this study. This work would not have been possible without their support and the dedication of the research community. Also, I would like to thank Faouzi and Mohib for placing their trust in me to entrust me with this project.

References

1. UMAP documentation. [Online]. Available: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

2. Deep learning for sensor-based activity recognition: A survey, by : Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, Lisha Hu

3. Sensor-based and vision-based human activity recognition: A comprehensive survey, by : L. Minh Dang, Kyungbok Min, Hanxiang Wang, Md. Jalil Piran, Cheol Hee Lee, Hyeonjoon Moon

4. Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges, and Opportunities, by : KAIXUAN CHEN, DALIN ZHANG, LINA YAO, BIN GUO, ZHIWEN YU, YUNHAO LIU

5. A comprehensive survey on gait analysis: History, parameters, approaches, pose estimation, and future work, by : Dimple Sethi, Sourabh Bharti, Chandra Prakash

6. Human Gait Analysis in Neurodegenerative Diseases: A Review, by : Grazia Cicirelli, Donato Impedovo, Vincenzo Dentamaro, Roberto Marani, Giuseppe Pirlo, Tiziana R. D'Orazio