Data augmentation study for rare diseases assessment with Deep Learning: Confocal Imaging analysis of Congenital Muscular Dystrophy

M. Frías[1], C. Jiménez-Mallebrera[2,3,4], C. Badosa[2], J.M. Porta*[5], M. Roldán *[1]

[1] Unitat de Microscòpia Confocal i Imatge Cel·lular, Departament de Medicina Genètica i Molecular, Institut Pediàtric de Malalties Rares, Hospital Sant Joan de Déu, Esplugues de Llobregat, Barcelona, Spain. marcos.frias@sjd.es, monica.roldan@sjd.es

[2] Laboratorio de Investigación Aplicada en Enfermedades Neuromusculares, Unidad de Patología Neuromuscular, Servicio de Neuropediatría, Institut de Recerca Sant Joan de Déu, Esplugues de Llobregat, Barcelona, Spain. cecilia.jimenez@sjd.es, mariacarmen.badosa@sjd.es

[3] Centro de Investigaciones Biomédicas en Red de Enfermedades Raras (CIBERER), Madrid, Spain.

[4] Departamento de Genética, Microbiología y Estadística, Universitat de Barcelona, Barcelona, Spain.

[5] Institut de Robòtica i Informàtica Industrial, UPC-CSIC, Barcelona, Spain. porta@iri.upc.edu.

*Both authors contributed equally to this work

Summary

Artificial Intelligence (AI) algorithms are widely used in healthcare nowadays. However, there are still fields where the application of these technologies could be challenging, such as rare diseases. In these cases, the main challenge arises from the reduced size of the available data sets. This paper proposes a data augmentation pipeline to address this challenge when using a Deep Learning (DL) algorithm to assess fibroblast cultures from skin biopsies to diagnose Collagen VI-related Congenital Muscular Dystrophy (COL6-CMD). Different data augmentation schemes are described in the literature. However, they must be used cautiously since they might result in overfitting. The results presented in this paper demonstrate that the right combination of data augmentation techniques results in a high diagnostic accuracy (up to 75.35% for the best approach) even with a scarce amount of data.

1. Introduction

Collagen VI-related Congenital Muscular Dystrophy (COL6-CMD) is a rare disease[1] that produces deficiencies in the structure of the protein Collagen VI [1]. Depending on the severity, different manifestations can be found. Together with the low prevalence and the typically limited understanding of rare diseases, it results very hard to achieve the correct diagnosis [2]. The visual analysis of fibroblast cultures with optical and photonic microscopy technologies is one of the most relevant techniques to reach the precise diagnosis of this disease [3]. More specifically, confocal microscopy provides a high signal sensibility able of extracting detailed information even in intermediate and mild manifestations of the disease [4]. Microscopy images present Collagen VI networks such as the ones shown in Figure 1. Different aspects are considered to assess them, such as the collagen fibres distribution and intensity or the nuclei shape, which might appear altered in pathological cases. Nevertheless, they are always inspected from the subjective point of view of the professional. For that reason, novel Artificial Intelligence (AI) techniques can provide more efficient and objective procedures to perform this task.

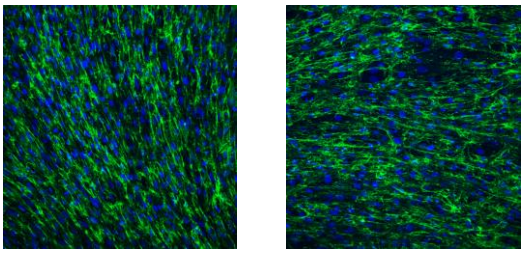

Figure 1 *Example of control (left) and patient (right) confocal images of primary fibroblasts cultures. The collagen fibres are shown in green and the cell nuclei in blue*

Some studies have demonstrated the viability of implementing AI approaches for microscopy samples and rare disease diagnosis [5]. However, the reduced number of patients suffering from these types of diseases generates a clear limitation regarding the amount of data available, which represents the main difficulty when using these technologies. This is especially true in Deep Learning (DL), where samples must be as representative as possible so that the algorithm can determine the key features to correctly perform the task [6]. Thus, there is a critical demand for novel data augmentation approaches in this context.

A few AI techniques have been applied to help on automatizing and accelerating the diagnosis process more quantitatively and objectively In COL6-CMD. In 2018,

[1] In Europe, a disease is considered as rare when its prevalence is below 1 in 2000.

ISBN: 978-84-17853-76-1

Bazaga et al. developed the first Computer-Aided Diagnosis (CAD) system for COL6-CMD based on a handcrafted DL algorithm trained with a small database formed by confocal microscopy acquisitions of primary fibroblast cultures extracted from skin biopsies. The homogeneous nature of the samples allowed for the generation of small, fixed-size image patches to generate a larger dataset without affecting the biological structure for identifying the disease. However, the technique implications in the algorithm performance were not analysed in detail [7].

This paper explores which are the key factors to consider when implementing a data augmentation pipeline to use DL technologies in a rare disease context. The COL6-CMD is analysed as an example in a limited-available dataset scenario. Three different data augmentation techniques are explored. First, the patches generation approach from Bazaga et al. is extended by trying different patch sizes. Additionally, different degrees of overlap between patches are considered. Finally, standard data augmentation transformations are applied.

2. Methods

2.1. Dataset

The amount of available data is crucial for an AI model to succeed. In 2018, the Neuromuscular Unit at the Research Institute of the Hospital Sant Joan de Déu together with the Confocal Microscopy and Cellular Imaging Unit started a database with fibroblasts culture confocal images acquired with a Leica TCS SP8 equipped with a white light laser and hybrid spectral detectors (Leica Microsystems GmbH, Mannheim, Germany). Confocal images, each consisting of 1024 × 1024 pixels, was accomplished using an HCX PL APO 20x/0.75 dry objective, with the confocal pinhole set to 1 Airy unit. Collagen VI fluorescence was excited by an argon laser (488 nm) and detected within the 500–560 nm range. Simultaneously, nuclei were excited with a blue diode laser (405 nm) and detected within the 420–460 nm range. To eliminate any potential channel interference, we employed sequential acquisition settings. Different images, called fields, are extracted from a single fibroblasts' culture. For each field, a series of ten sections, spaced 1.5 μm apart along the focal axis (Z-stack), were acquired. These individual sections were subsequently integrated into a maximum intensity projection, resulting in a single comprehensive image.

In 2022, this dataset contained 411 TIFF images, 221 were labelled as controls and 190 as patients. They are extracted from 24 control biopsies and 19 patient biopsies. This images set is divided in training and testing subsets, with an 80-20% ratio respectively. It is crucial to remark that there is a dependency between the different fields extracted from the same cells' culture. Thus, they cannot be considered as independent (i.e., all the fields from the same cells' culture can only be assigned to either the training or the test set, but not split between them).

2.2. Data augmentation

The dataset size was not enough for training a Convolutional Neural Network (CNN) satisfactorily. For that reason, a data augmentation pipeline was developed. First, each image was split into small patches. In that way, the different patches could be classified individually to be posteriorly grouped at image level following a majority voting system. Different patch sizes were explored: 64x64x, 128x128 and 224x224 pixels. Moreover, partial overlapping was considered to further increase the dataset dimension. The degree of overlap considered were 0%, 25%, 50%, and 75%. Table 1 shows the resulting number of patches generated when applying the different combinations of data augmentation. Additionally, other transformations were performed on-the-fly. Such transformations are random rotations and horizontal and vertical flips, which were applied each time data was set into the network with a 0.5 probability.

		Patch size (in pixels)		
		64x64	**128x128**	**224x224**
Overlap	**0%**	105.216	26.304	10.275
	25%	180.840	49.320	14.385
	50%	394.560	92.064	32.880
	75%	1.528.920	345.240	104.805

Table 1 *Number of total patches for the different patch sizes and the degree of overlap between patches, after manual data augmentation*

2.3. Convolutional Neural Network

To classify the available images, we used the EfficientNetB0 network since it outperforms other alternatives with a relatively small number of parameters [8]. The network is initialized with pre-trained weights that are latter fine-tuned for the considered problem [9]. In this way, 5 epochs (with a batch size of 64) are enough for the training algorithm to converge.

The reduced dataset size in these type of pathologies causes the model to be very dependent on the split between training and testing. For that reason, we follow the K-fold cross-validation technique: the results are averaged over five different training-test partitions [10]. Moreover, an accumulative confusion matrix for the five partitions is compute to generate more robust results.

3. Results

Both the training and test accuracies and the Area Under the Curve (AUC) were used quantify the model performance when applying different data augmentation techniques. Tables 2 and 3 summarize the obtained results.

ISBN: 978-84-17853-76-1

			Patch size (in pixels)		
			64x64	128x128	224x224
Overlap	0%	Train	95.44 ± 0.27	97.59 ± 0.24	97.91 ± 0.12
		Test	60.40 ± 2.08	62.54 ± 4.91	**74.53 ± 3.79**
	25%	Train	97.00 ± 0.16	97.96 ± 0.15	98.52 ± 0.14
		Test	60.04 ± 4.48	67.12 ± 6.78	73.31 ± 4.86
	50%	Train	98.25 ± 0.18	99.12 ± 0.18	99.17 ± 0.04
		Test	57.28 ± 5.26	64.57 ± 5.47	72.77 ± 4.76
	75%	Train	99.32 ± 0.03	99.64 ± 0.01	99.61 ± 0.03
		Test	55.83 ± 4.62	66.8 ± 3.08	72.60 ± 2.94

Table 2 *Average train and test accuracies obtained with each combination of patch size and degree of patch overlap over the 5-fold cross-validation. The standard error is also shown.*

		Patch size (in pixels)		
		64x64	128x128	224x224
Overlap	0%	0.671 ± 0.036	0.659 ± 0.078	**0.811 ± 0.041**
	25%	0.612 ± 0.061	0.706 ± 0.105	0.809 ± 0.054
	50%	0.546 ± 0.092	0.733 ± 0.063	0.767 ± 0.061
	75%	0.562 ± 0.066	0.705 ± 0.047	0.792 ± 0.042

Table 3 *Average AUC obtained with each combination of patch size and degree of patch overlap over the 5-fold cross-validation. The standard error is also shown.*

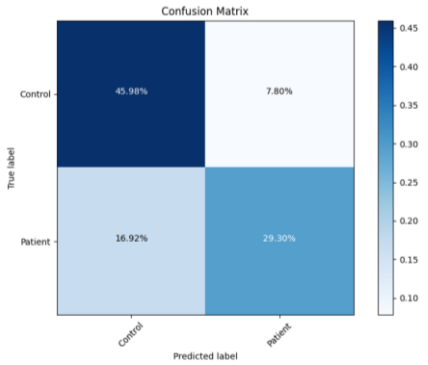

Figure 2 *Accumulative confusion matrix at patch level obtained with patches of 224x224 pixels, with no overlap between patches, and with standard data augmentation techniques applied on-the-fly.*

4. Discussion

The implementation of DL techniques in healthcare is flourishing, largely due to the increasingly available medical data and to the more powerful computer resources. The main factor for a DL algorithm to succeed is to have enough available data. Otherwise, the learned model would not properly generalize the desired task. For this reason, the application of this technology to a rare disease, where data is severely limited, is a very challenging task that requires the development of data augmentation techniques [11].

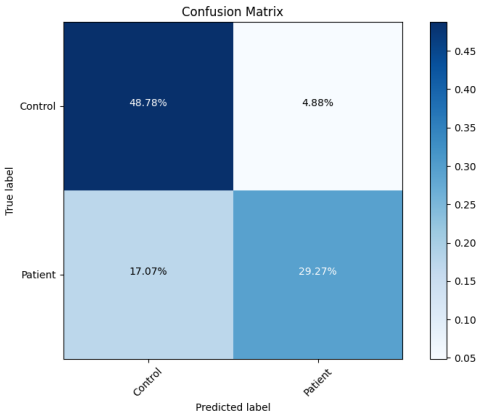

Figure 3 *Accumulative confusion matrix at fibroblasts' culture level obtained after majority voting of patches of 224x224 pixels, with no overlap between patches, and with standard data augmentation techniques applied on-the-fly.*

Different approaches have been explored in this paper to deal with data limitations in a COL6-CMD clinical scenario. This disease presents alterations over a homogeneous collagen network, which allowed the generation of smaller patches to increase the number of samples, while preserving the complete biological structure [12]. However, the balance between the size and the overlap of these patches is key for the technique to succeed, since very small regions might not include enough biological information while big patches might not generate enough samples for effectively training the network.

Results showed that it is possible to achieve a good performance when implementing a DL algorithm if the adequate data augmentation techniques are applied. The best results (with a 74.53% of testing accuracy) were achieved with the 224x224 patch size with no overlap. The results indicate that the smaller the patch, the worst the result. Regarding the overlap between patches, in general, no improvement appears when increasing the percentage of overlap. The higher the overlap, the larger the data set, but also the larger the possibility of overfitting the model, since the network is trained repetitively with the same sub-patches. This causes the testing metrics to decrease while the training metrics become higher [13]. This behaviour can be observed in Table 2, which suggests that overlapping patches might not be a relevant data augmentation technique for this type of biological samples and clinical context. The high AUC values confirm the capability of the model to correctly discern between classes.

Standard on-the-fly data augmentation techniques in DL slightly improve the model performance, as expected [14].

Control samples are more coherent between them, but patient samples are more diverse since the disease can appear in different manifestations and degrees of severity. Therefore, not both classes are classified equally correctly. The algorithm learnt to recognize patches coming from control cultures almost perfectly since only

ISBN: 978-84-17853-76-1

14% of the total healthy patches were misclassified over the 5-fold cross-validation process. On the contrary, the patches extracted from patient biopsies are more complex to identify. In this case, 37% of the total patches were misclassified as control. At the biopsy level, using the majority voting approach, similar results were obtained. This confirms the effective extrapolation of this data augmentation technique at the person level, in order to give an indicative diagnosis to the person suffering from this disease.

Alternative data augmentation techniques could be explored to complement this pipeline. One option could be to explore generative models such as Generative Adversarial Networks (GANs), which aim to learn to generate synthetic realistic images to enlarge the database. This is a complex task prone to mode collapse and the production of hallucinations. Nevertheless, it is a very promising field of work in which many advances have been recently done [15]. Additionally, shared protocols between different centres could be established to allow for sharing their data and increment the size of the image data sets.

In conclusion, many aspects have to be considered for developing an effective data augmentation pipeline when applying DL techniques to a rare disease. The dataset size and organization, the knowledge about the underlying biologic structure, and the photonic techniques used for the image acquisitions requires professionals from multiple fields to work together. In this way, the challenges appearing when using DL techniques on the small datasets available in the context of rare and ultra-rare diseases can be successfully overcome.

Acknowledgments

Funded by the European Union. We are indebted to the HORIZON–MSCA–2022–DN, Improving BiomEdical diagnosis through LIGHT-based technologies and machine learning “BE-LIGHT” (GA nº 101119924 – BE-LIGHT).

References

[1] Natera-de Benito D et al. Association of Initial Maximal Motor Ability With Long-term Functional Outcome in Patients With COL6-Related Dystrophies. Neurology. 2021 Mar 9;96(10):e1413-e1424.

[2] Wakap SN; Lambert DM; Olry A; Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 28:165–173. (2020).

[3] Jimenez-Mallebrera, C., Maioli, M. A., Kim, J., Brown, S. C., Feng, L., Lampe, A. K., Bushby, K., Hicks, D., Flanigan, K. M., Bönnemann, C. G., Sewry, C. A., & Muntoni, F. (2006). A comparative analysis of collagen VI production in muscle, skin and fibroblasts from 14 Ullrich Congenital Muscular Dystrophy patients with dominant and recessive COL6A mutations. Neuromuscular Disorders, 16(9-10), 571-582. https://doi.org/10.1016/j.nmd.2006.07.015

[4] Nwaneshiudu, A., Kuschal, C., Sakamoto, F. H., Rox Anderson, R., Schwarzenberger, K., & Young, R. C. (2012). Introduction to confocal microscopy. Journal of Investigative Dermatology, 132(12), 1-5. https://doi.org/10.1038/jid.2012.429

[5] Lee, J. H., Liu, C., Kim, J., Chen, Z., Sun, Y., Rogers, J. R., Chung, W. K., & Weng, C. (2022). Deep Learning for Rare Disease: A scoping review. Journal of Biomedical Informatics, 135, 104227. https://doi.org/10.1016/j.jbi.2022.104227

[6] Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z

[7] Bazaga, A., Roldán, M., Badosa, C., Jimenez-Mallebrera, C., & Porta, J. M. (2019). A convolutional neural network for the automatic diagnosis of collagen VI-related muscular dystrophies. Applied Soft Computing, 85, 105772. https://doi.org/10.1016/j.asoc.2019.105772

[8] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR

[9] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A Survey on Deep Transfer Learning. Lecture Notes in Computer Science, 270-279. https://doi.org/10.1007/978-3-030-01424-7_27

[10] Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (2012, April). The'K'in K-fold Cross Validation. In ESANN (pp. 441-446).

[11] Alsentzer, E., Li, M. M., Kobren, S. N., Undiagnosed Diseases Network, Kohane, I. S., & Zitnik, M. (2022). Deep learning for diagnosing patients with rare genetic diseases. medRxiv, 2022-12.

[12] Garcea, F., Serra, A., Lamberti, F., & Morra, L. (2023). Data Augmentation for Medical Imaging: A Systematic Literature Review. Computers in Biology and Medicine, 152, 106391. https://doi.org/10.1016/j.compbiomed.2022.106391

[13] Ying, X. (2019, February). An overview of overfitting and its solutions. In Journal of physics: Conference series (Vol. 1168, p. 022022). IOP Publishing.

[14] Chlap, P., Huang, M., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology, 65(5), 545-563. https://doi.org/10.1111/1754-9485.13261

[15] Zhao, S., Liu, Z., Lin, J., Zhu, J. Y., & Han, S. (2020). Differentiable augmentation for data-efficient gan training. Advances in neural information processing systems, 33, 7559-7570.