Advancing Diagnosis of COL6-CMD: AI and Transfer Learning in Muscle Tissue Analysis

I. Georvasilis[1], M. Frías[1], C. Badosa[2], A. López-Márquez[2,3,4], C. Jiménez-Mallebrera[2,3,4], 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.*
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.*
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.*

Contact mail: ioannis.georvasilis@sjd.es, monica.roldan@sjd.es

Funded by the European Union

INTRODUCTION

Diagnosing rare **neuromuscular disorders**, such as **Collagen VI-related Congenital Muscular Dystrophy (COL6-RD)**, presents significant challenges due to its **low prevalence** and the **broad spectrum of clinical presentations**, which often overlap with other forms of muscular dystrophy. Although genetic testing is the definitive diagnostic method, it is time-consuming and not always immediately available. A common diagnostic approach involves **immunofluorescence analysis** to evaluate **collagen VI expression patterns**. However, this method is laborious, subjective, and depends heavily on the clinician's expertise.

This study aims to develop a **deep learning-based pipeline** to enhance the diagnostic process for COL6-RD by leveraging artificial intelligence (AI) to classify collagen VI expression patterns more objectively and efficiently. To achieve this, we used a dataset derived from a mouse model of COL6-RD, including 154 confocal microscopy images of muscular tissue categorized into three genotypic groups: **wild-type (WT)**, **heterozygous (HT)**, and **homozygous (HOM)**. This model replicates key aspects of the disease, providing a controlled framework for testing AI-based diagnostic approaches.

The incorporation of AI offers promising potential to streamline and standardize the diagnostic process for rare diseases like COL6-RD. However, effective AI algorithms require large and representative datasets to capture the full heterogeneity of clinical presentations. Our approach addresses this challenge by applying **transfer learning** with **Cellpose**, a pretrained model for **cellular segmentation**, **fine-tuned** with our dataset to improve classification accuracy and overcome the limitations of data scarcity in rare diseases.

MATERIALS & METHODS

COLLAGEN VI

- **Role in ECM**: It is a key Extracellular Matrix (ECM) protein forming microfibrils, that maintain muscle integrity and support cellular adhesion.
- **Genetic Variants and Mutations**:
  - **WT** (Wild-Type) - Normal collagen production.
  - **HET** (Heterozygotous) – one mutated allele.
  - **HOM** (Homozygous) – two mutated alleles.

BIOPSY SIGNIFICANCE

- **Complementing Genetic Testing**: Critical for resolving inconclusive genetic diagnoses.
- **Heterozygous Insights:** Demonstrates reduced allele expression, correlating with milder symptoms.

COL6-CMD

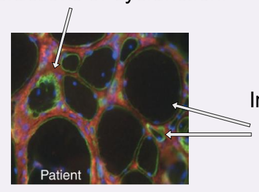
- **Prelevance**: <1 in 100.000
- **Disease Course**: Severity and progression vary significantly by subtype.
- **Treatment**: No cure treatment available; treatment focuses on symptom management.

ANIMAL MODEL

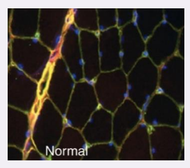
- **Stimulating Disease**: Replicates muscle responses across genotypic variants.
- **Spectrum Representation**: Captures full disease spectrum, including intermediate cases

COLOCALIZATION

In patient biopsies, colocalization is disrupted, indicating altered protein interactions or dysfunction.



SHAPE VARIABILITY

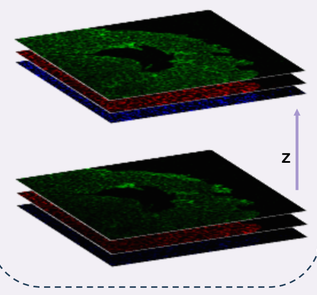


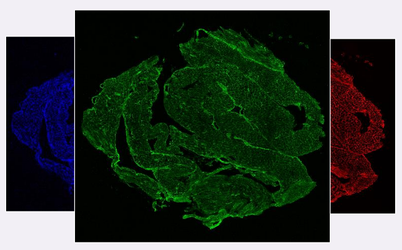
Irregular cell shapes reflect cellular dysfunction and tissue damage.

INSPECTED MUSCLES

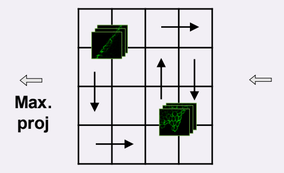
- Biceps (BI)
- Diaphragm (DIA)
- Quadriceps (QUA)
- Gastrocnemius (GSN)
- Soleus (SOL)
- Tibialis anterior (TA)
- Triceps (TRI)

Each section is composed of consistently acquired, independent **3D volumetric fields** across 3 channels: **Nuclei**, **Perlecan**, **COL6**.





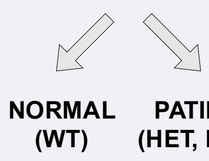
154 TISSUE BIOPSIES
50 WT - 52 HET - 52 HOM



All independent **fields** are then mapped together into a complete spatial 2D representation of the tissue.

RESULTS

BINARY CLASSIFICATION



TRAINING DETAILS

- Epochs: **100**
- Optimizer: **Adam**
- Loss: **Cross Entropy**
- **K-fold Cross Validation**

MODIFICATIONS

- **Cellpose** requires **2-channel input**; we adapted it for **3-channel input**, including **COL6**.
- **Solution:** Duplicated **ECM** weight to accommodate the additional channel.

DATASET REFINEMENT PROCESS

- Patching strategy generated over 50K total patches.
- **Low intensity-variation patches** were automatically excluded.
- Patches with misoriented fibers and vessels were manualy discarded.
- **Final dataset: 5K** remaining patches for training.

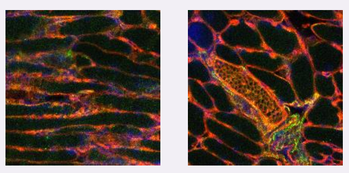


Figure 1. *Images depicting discarded misoriented fibers (left) and incorrectly aligned vessels (right).*

Model	Accuracy	F1- Score	AUC
Base	0.79	0.79	0.79
Cellpose	**0.82**	**0.82**	**0.86**
Base*	0.68	0.68	0.68
Cellpose*	0.74	0.75	0.78

Table 1. *Performance comparison between the pretrained (Cellpose) and randomly initialized (Base) model, after 5* and 100 epochs of training. The pretrained model demonstrates significantly faster convergence, greater accuracy, and superior overall performance compared to the base one.*

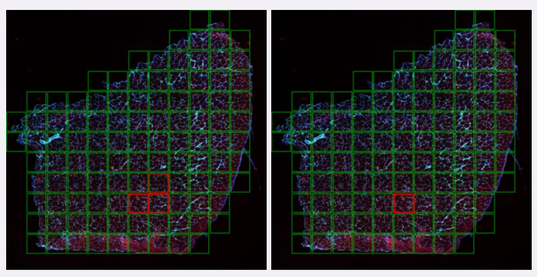
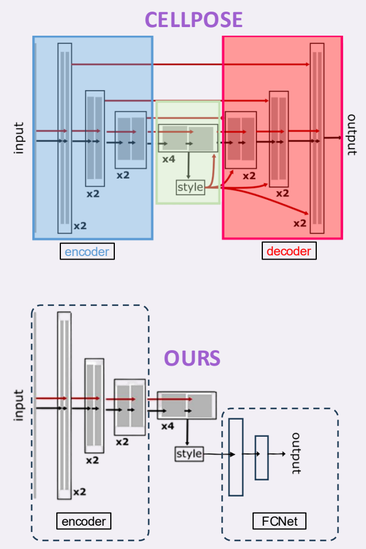


Figure 2. *Tissue image showing patch-wise predictions from Cellpose (right) and Base (left). Green rectangles indicate correct predictions (WT), while red rectangles highlight incorrect ones. The pretrained model shows fewer errors. Majority voting achieves 96% accuracy, providing additional insight for the clinicians.*



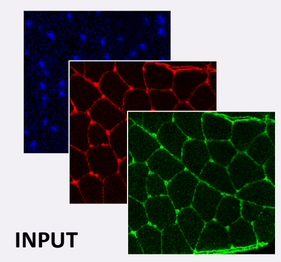
Automated Segmentation to Measure Variability
- Speed and Adaptability: Faster segmentation across diverse cell morphologies and collagen intensity levels.

Cellpose Overview
- CNN for automated cell segmentation. Trained on a variety of cell types and imaging conditions.

Transfer Learning Approach
- Knowledge transfer from a pre-trained model (**Cellpose**) on a new, related task (**Classification**).
- **Fine-tuned** on small, specialized dataset typical in rare disease research.

PATCHING STRATEGY

- **Focus on Local Features**: Isolates relevant areas like muscle fibers and nuclei.
- **Computational Efficiency**: Resolution too high for direct processing by Neural Networks.
- **Enhances Data Augmentation**: Increases sample size for better model generalization.



CONCLUSIONS

- We successfully demonstrated the power of AI and Transfer Learning in enhancing generalization for rare disease diagnoses, such as COL6-CMD, even with limited datasets.
- Leveraging pre-trained neural networks enables faster convergence, making it particularly beneficial for cost-efficient applications.
- The Patching Strategy enhances interpretability by highlighting regional predictions within the entire tissue sample.
- The model exhibits challenges in accurately distinguishing the HET class, which represents intermediate cases.

FUTURE WORK

- **Database Expansion**: Increase the dataset with additional diverse images for robustness
- **3D Image Training**: Apply Transfer Learning to full-volume 3D images instead of maximum projections to capture depth and comprehensive structural information.
- **Refining the Pre-training Strategy:** Explore alternative pre-trained models beyond Cellpose and incorporate self-supervised pre-training approaches to enhance performance.
- **Enhancing the Patching Strategy**: Focus on localizing nuclei and defining their spatial neighborhood as the basis for extracted patches.

REFERENCES

1. 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.
2. Schaefer, J., Lehne, M., Schepers, J., Prasser, F., Thun, S., 2020. The use of machine learning in rare diseases: a scoping review. Orphanet Journal of Rare Diseases 15. doi:10.1186/s13023-020-01424-6.
3. Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose: A generalist algorithm for cellular segmentation. Nature Methods, 18(1), 100-106. https://doi.org/10.1038/s41592-020-01018-x.
4. Briguet, A., Courdier-Fruh, I., Foster, M., Meier, T., & Magyar, J. P. (2004). Histological parameters for the quantitative assessment of muscular dystrophy in the mdx-mouse. Neuromuscular Disorders, 14(10), 675-682. https://doi.org/10.1016/j.nmd.2004.06.008.