Unsupervised machine learning methods to uncover multiple sclerosis subtypes on MRI-derived data

Thesis by

Anna Gabriela Vargas Fichera

In Fulfillment of the Requirements for the degree
of
MSc. in Bioinformatics

AARHUS UNIVERSITY
Natural & Technical Sciences

Department of Molecular Biology and Genetics

Defense: June 27, 2025

Thesis submitted: June 15, 2025

Main Supervisor: Assoc. Prof. Simon Fristed Eskildsen
Department of Clinical Medicine - Center of Functionally Integrative Neuroscience, Aarhus University

Co-Supervisor: Assoc. Prof. Palle Villesen
Department of Clinical Medicine - Unit of Bioinformatics, Aarhus University

Acknowledgments

I would like to express my gratitude to Assoc. Prof. Simon Fristed Eskildsen, for giving me the opportunity to work in his research group, and for his guidance and expertise. I am also grateful to Postdoc Mikkel Karl Emil Nygaard, for his valuable advice, knowledge, and guidance throughout this project. I would also like to thank Assoc. Prof. Palle Villensen, for his aid and contributions.

Abstract

Multiple sclerosis (MS) is a prevalent neurodegenerative disorder characterized by heterogeneous pathology and clinical progression. Nowadays, MS is classified into four clinical subtypes. However, the biological basis and the underlying mechanisms that differentiate each subtype remain unknown. Recent advances in neuroimaging and machine learning offer the potential to uncover hidden disease structures. However, it is still unclear whether Magnetic Resonance Imaging (MRI)-derived features can meaningfully distinguish biological subtypes of MS. In the present study, a multicohort dataset with MRI-derived data, consisiting of diffusion and volumetric features, from three different studies was used with the aim of uncovering possible MS subtypes using unsupervised machine learning. Dimensionality reduction techniques consistently revealed clustering by cohort, especially in diffusion-derived metrics, suggesting a strong batch effect rather than an underlying disease structure. Using different batch correction strategies, the cohort-specific variance was reduced, but no biological subtypes were identified. Further association analysis revealed that disease progression features could have explained the cohort-driven differences. Before batch correction, the clinical variables were mostly associated to mean diffusivity features, reflecting early pathological processes. After batch correction, fractional anisotropy and volumetric features became more prominent, potentially indicating later structural changes. These findings do not represent different subtypes of the disease, rather disease progression and biological patterns from the different cohorts. Further, even if an association was found between clinical and biological features, the presence of an unexplained technical variance should not be discarded.

List of Abbreviations

AI: Artificial Intelligence

ANOVA: Analysis of Variance

AveDist: Average Euclidean distance

CIS: Clinically Isolated Syndrome

CNS: Central Nervous System

CSF: Cerebrospinal Fluid

DTI: Diffusion Tensor Imaging

DKI: Diffusion Kurtosis Imaging

EDSS: Expanded Disability Status Scale

EMSES: Early Exercise Efforts in Multiple Sclerosis

EXBRAIN: Aerobic Exercise and Brain Health in Multiple Sclerosis

FA: Fractional Anisotropy

FDR: False Discovery Rate

GM: Grey Matter

ICC: Intracranial volume

MD: Mean Diffusivity

MANOVA: Multivariate Analysis of Variance

ML: Machine Learning

MRI: Magnetic Resonance Imaging

MS: Multiple Sclerosis

NAWM: Normal-appearing white matter

NfL: Neurofilament light chain marker

PASAT: Paced Auditory Serial Addition Test

PCA: Principal Components Analysis

PoTOMS: Power Training in Older Multiple Sclerosis Patients

PPMS: Primary Progressive Multiple Sclerosis

RRMS: Relapsing-Remitting Multiple Sclerosis

RT: Reaction Times

SDMT: Symbol Digit Modalities Test

SixMWT: Six Minute Walk Test

SPMS: Secondary Progressive Multiple Sclerosis

SRT: Selective Reminding Test

SSST: Six Spot Step Test

SuStaIn: Subtype and Stage Inference

TP: Timepoint

t-SNE: t-distributed stochastic neighbor embedding

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

WM: White Matter

Contents