Impact of Data Patterns on Biotype identification Using Machine Learning

Yuetong (Jordan) Yu[1,2], Ruiyang Ge[1,2], Ilker Hacihaliloglu[1,3,4], Alexander Rauscher[1,5], Roger Tam[1,6], Sophia Frangou[1,2,7]

[1]Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Canada [2]Department of Psychiatry, University of British Columbia, Canada [3]Department of Radiology, University of British Columbia, Canada [4]Department of Medicine, University of British Columbia, Canada [5]Department of Pediatrics, University of British Columbia, Canada [6]School of Biomedical Engineering, University of British Columbia, Canada [7]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, USA

Background

- Machine learning algorithms have been widely applied for identifying patient subtypes, results have been inconsistent across studies.
- These inconsistencies are often attributed to algorithmic limitations, yet an overlooked factor may be the statistical properties of the input data.

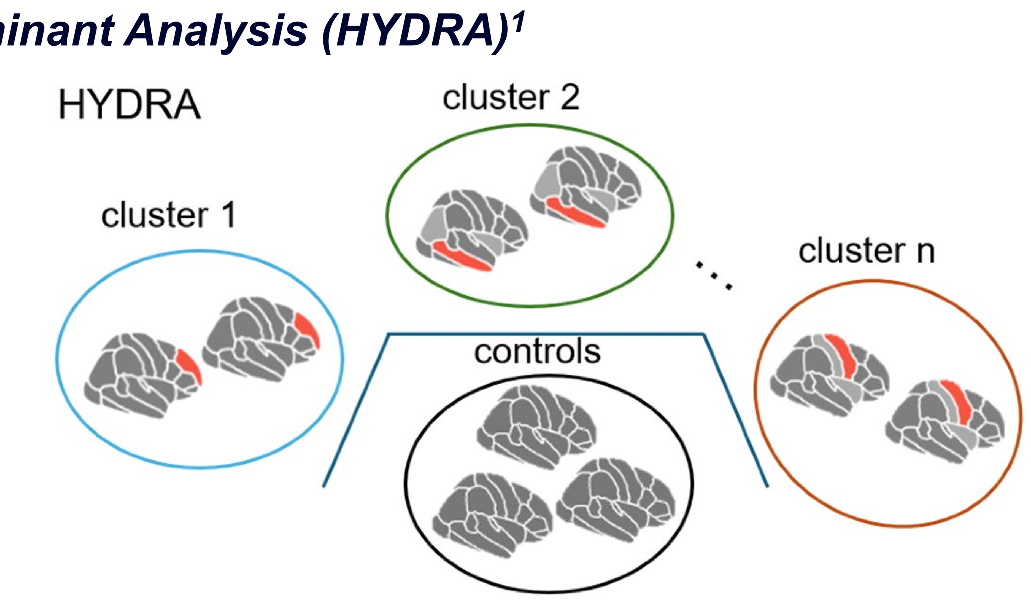
Objective

- To investigate the contribution of data patterns on algorithm performance by leveraging synthetic brain morphometry data as an exemplar

Methods

1. HeterogeneitY through DiscRiminant Analysis (HYDRA)[1]

- Semi-supervised, non-linear clustering algorithm integrating binary classification.
- Constructs a convex polytope using multiple linear classifiers to separate patient subgroups from controls.

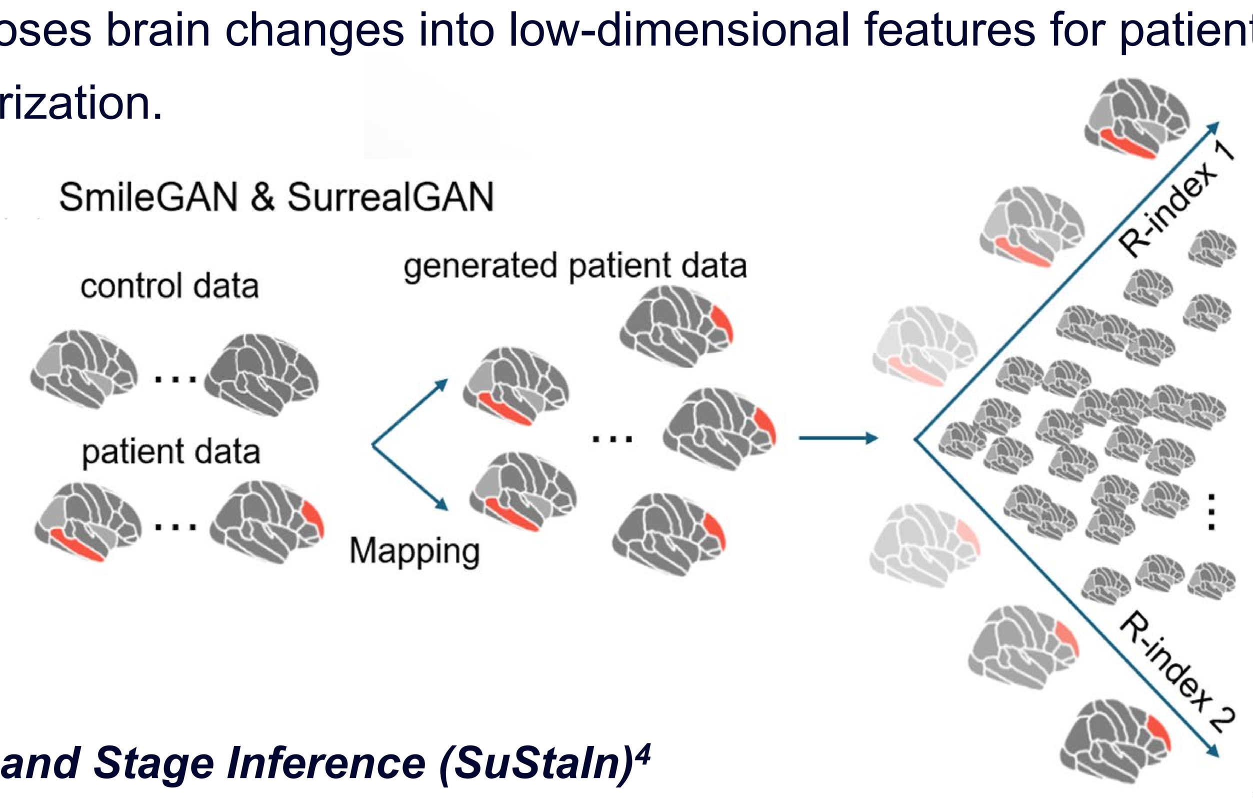


2. SeMI-supervised cLustEring via Generative Adversarial Network (SmileGAN)[2]

- GAN-based semi-supervised clustering approach.
- Learns patient-specific differences by generating pseudo-patient data from controls.

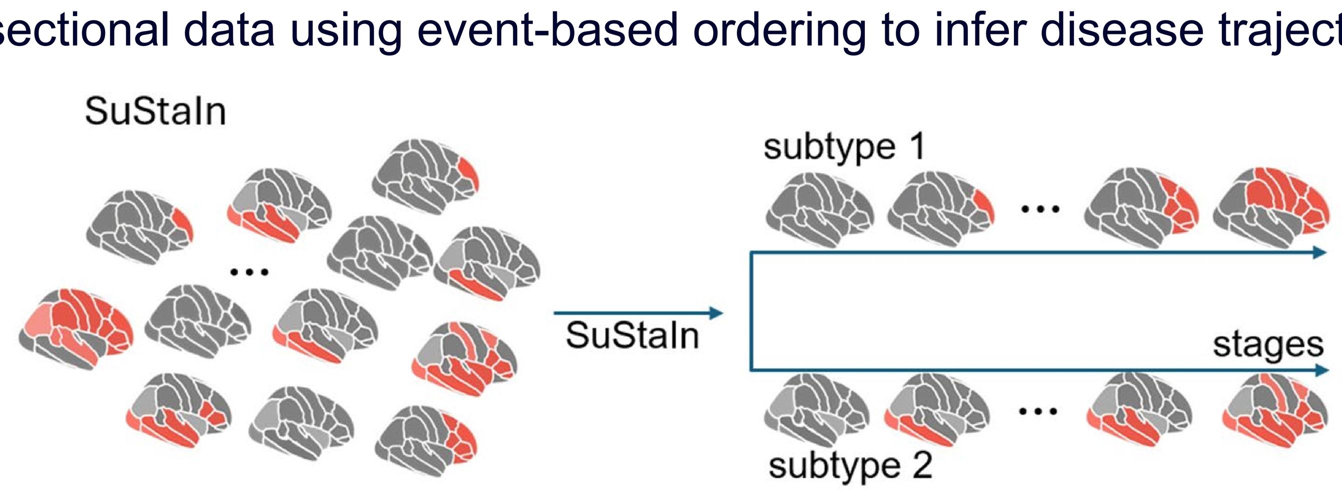
3. Semi-Supervised Representation Learning via GAN (SurrealGAN)[3]

- GAN-based representation learning method extracting R-indices that capture disease-related deviations.
- Decomposes brain changes into low-dimensional features for patient characterization.



4. Subtype and Stage Inference (SuStaIn)[4]

- Unsupervised clustering method identifying temporal progression patterns from cross-sectional data using event-based ordering to infer disease trajectories.



Data

We utilize five synthetic datasets (Syn1–Syn5) to systematically assess clustering algorithms by controlling data patterns.

- **Syn1 & Syn2**: Pseudo patient (Pseudo-PT) data forms three clusters. Syn1 includes 1,200 participants (600 controls, 600 Pseudo-PT), 145 variables. Syn2 contains 1,500 participants (600 controls, 900 Pseudo-PT), 100 variables.
- **Syn3–Syn5** are generated from real neuroimaging data (Human Connectome Project, n = 1,108), with 508 participants as a reference group and 600 Pseudo-PTs with 3 to 6 distinct clusters. Each participant contains 150 anatomical variables.

150 anatomical measures: FreeSurfer outputs

- **Syn3:** Monotonic increase in chosen variable values.
- **Syn4**: Monotonic decrease in chosen variable values.
- **Syn5**: Mixed increases and decreases.

Results

SuStaIn

- Unable to process >17 variables, limiting its applicability to synthetic datasets with >100 variables.
- Did not produce valid results due to the failure of convergence.

HYDRA

- Accuracy: 0.99 (Syn1), 0.52 (Syn2).
- Varied performance (0.45–0.99) across other datasets (Figure 2).
- Performance inconsistencies could not be fully explained by data distribution characteristics.

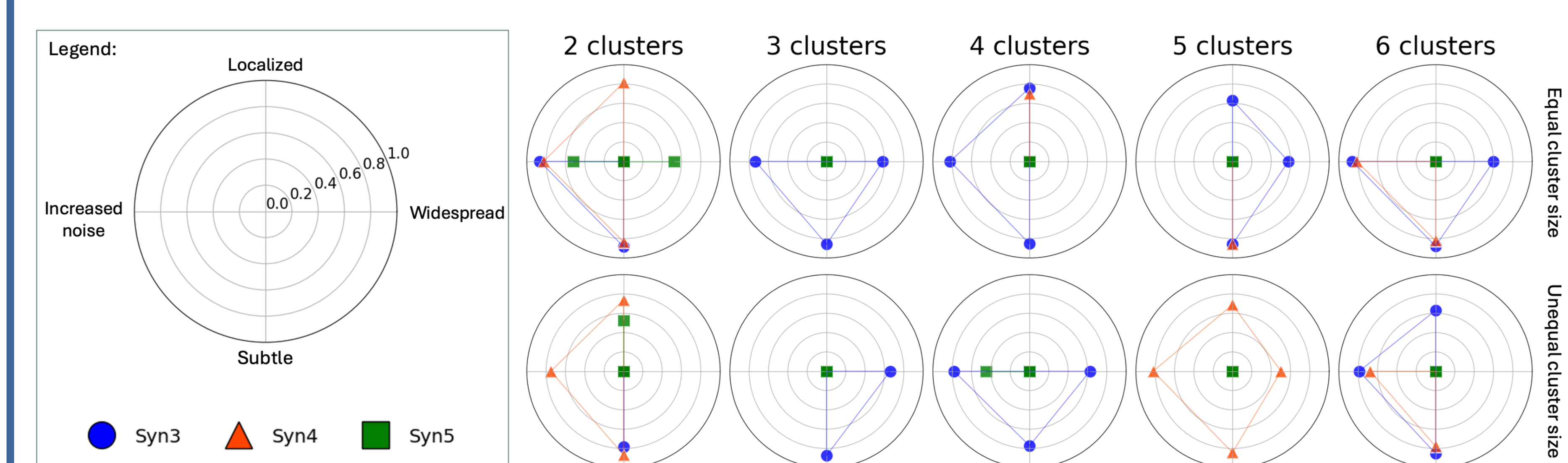


Figure 2. HYDRA Accuracy Across Synthetic Datasets Syn3-Syn5

SmileGAN

- Accuracy: 0.99 (Syn1), 0.61 (Syn2).
- Performance ranged from 0.37–0.99 (Figure 3).
- Performed best in Syn1 and Syn3 (2 Pseudo-PT clusters of equal size).
- Performance declined with increasing cluster complexity and variation in data changes.

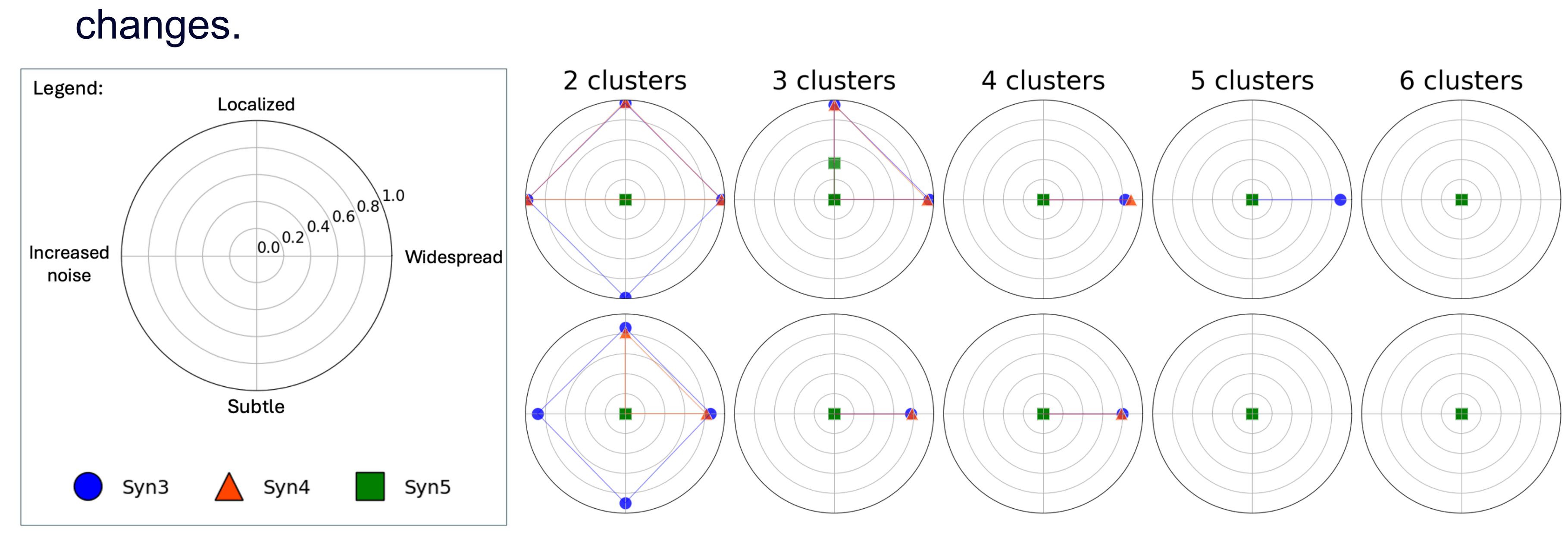


Figure 3. SmileGAN Accuracy Across Synthetic Datasets Syn3-Syn5

SurrealGAN

- Accuracy: 0.97 (Syn1), 0.83 (Syn2).
- Performance ranged from 0.38–0.99 (Figure 4).
- Achieved highest accuracy (0.99) in Syn3, performed well across variations but struggled with Syn5, especially in complex cluster settings.

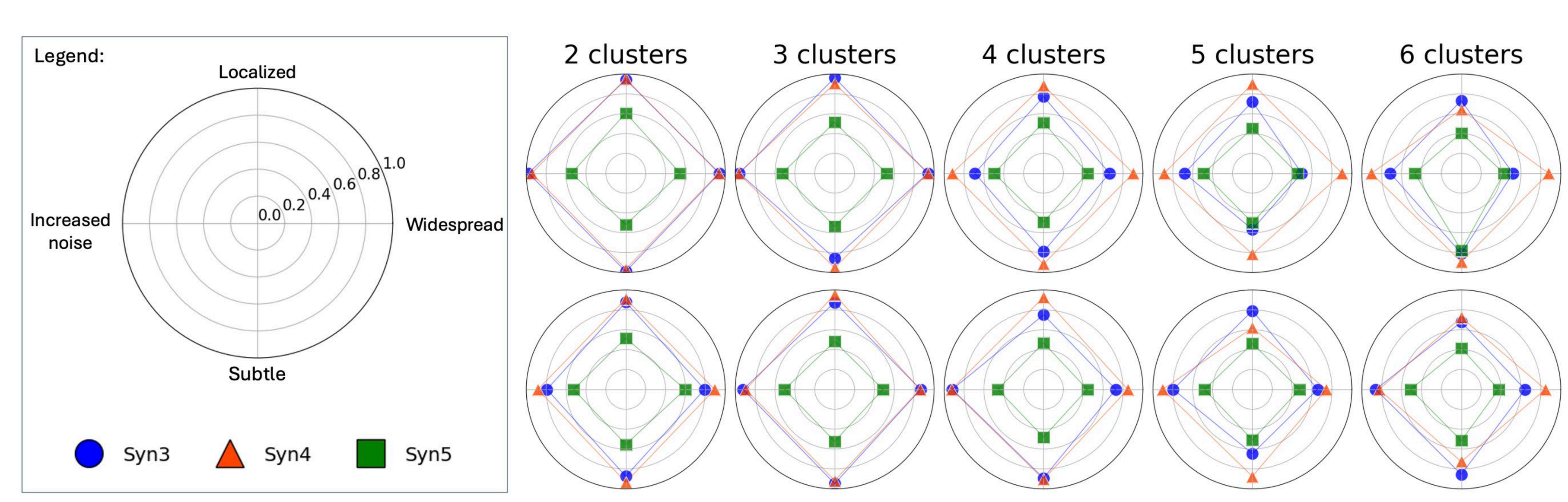


Figure 4: SurrealGAN Accuracy Across Synthetic Datasets Syn3-Syn5

Conclusion

- Dataset characteristics significantly influence algorithm performance, often more than algorithmic design.
- The findings emphasize the need for rigorous validation using synthetic data before real-world application.
- These insights extend beyond neuroimaging and have implications for machine learning applications in biomedical research.

Reference

1. Varol et al., NeuroImage, 2017. DOI: 10.1016/j.neuroimage.2016.02.041
2. Yang et al., Nature Communications, 2021. DOI: https://doi.org/10.1038/s41467-021-26703-z
3. Yang et al., ICLR 2022. DOI: https://doi.org/10.48550/arXiv.2205.04523
4. Young et al., Nature Communications, 2018. DOI: https://doi.org/10.1038/s41467-018-05892-0