NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

MICCAI2022 Singapore

SCAN ME

Semi-supervised learning with data harmonisation for biomarker discovery from resting state fMRI

Yi Hao Chan, Wei Chee Yew and Jagath C. Rajapakse

Introduction

Motivation: Datasets collected from a single institution contain valuable information about a specific cohort, but they are too small for existing approaches to generate useful biomarkers.

Problem: Ungeneralisable biomarkers, especially if they are from models fitted on small and high-dimensional datasets.

- Pooling multiple datasets together introduces issues such as site effects and labelling inconsistencies.
- Biomarkers from multi-site data are not always relevant to specific sites, but we lack tools to distinguish them.

Key Takeaways

1. SHRED separate site-specific and site-invariant biomarkers, allowing **cohort-specific** insights from **small datasets** to be created. SHRED also produces **personalised biomarkers**.
2. Our results on the ADHD-200 and ABIDE datasets shows the importance of accounting for both **site effects** and **labelling inconsistencies** in retrospectively aggregated datasets. With more neuroimaging research conducted on such datasets, data harmonisation and semi-supervised learning should be performed before biomarker generation.

Method

SHRED: an **interpretable** encoder-decoder-classifier architecture combining site difference removal with semi-supervised learning, trained end-to-end via a **novel multi-objective loss**.

Cross-Entropy

$$L_C = -\frac{1}{|S_l|}\sum_{j\in S_l}\sum_{c=1}^{C} y_{jc}\,\log(\hat{y}_{jc})$$

Likelihood loss

$$L_L = \frac{1}{2N}\sum_{j=1}^{N}\sum_{v=1}^{V}\log(2\pi\sigma_v^2) + \frac{(x_{jv}-\hat{x}_{jv})^2}{2\sigma_v^2}$$

KL Divergence

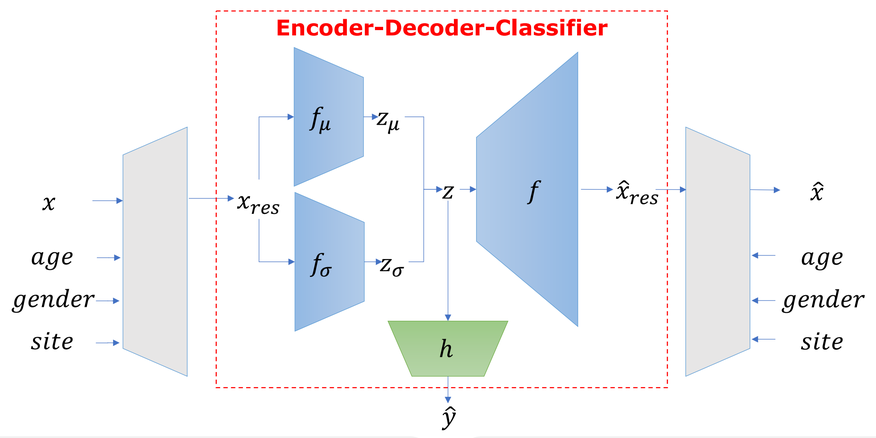
$$L_D = \frac{1}{2N}\sum_{j=1}^{N}\sum_{v=1}^{V} z_{\sigma_{jv}}^2 + z_{\mu_{jv}}^2 - 2\,\log(z_{\sigma_{jv}}) - 1$$

Reconstruction

$$L_R = \frac{1}{N}\sum_{j=1}^{N}\sum_{v=1}^{V}\varepsilon_{jv}^2 + (\alpha_v + M_{jv}\beta_v - \bar{x}_{jv})^2$$

Overall

$$L = L_C + \gamma_1 L_L + \gamma_2 L_D + \gamma_3 L_R$$



Harmonization Module

$$x_{res} = \frac{x_{ijv} - \alpha_v - M_{jv}\beta_v - \gamma_{iv}}{\delta_{iv}} + \alpha_v + M_{jv}\beta_v$$

Inverse Harmonization

$$\hat{x}_{ijv} = \alpha_v + M_{jv}\beta_v + \gamma_{iv} + \delta_{iv}(\hat{x}_{res} - \alpha_v - M_{jv}\beta_v)$$

Biomarker generation

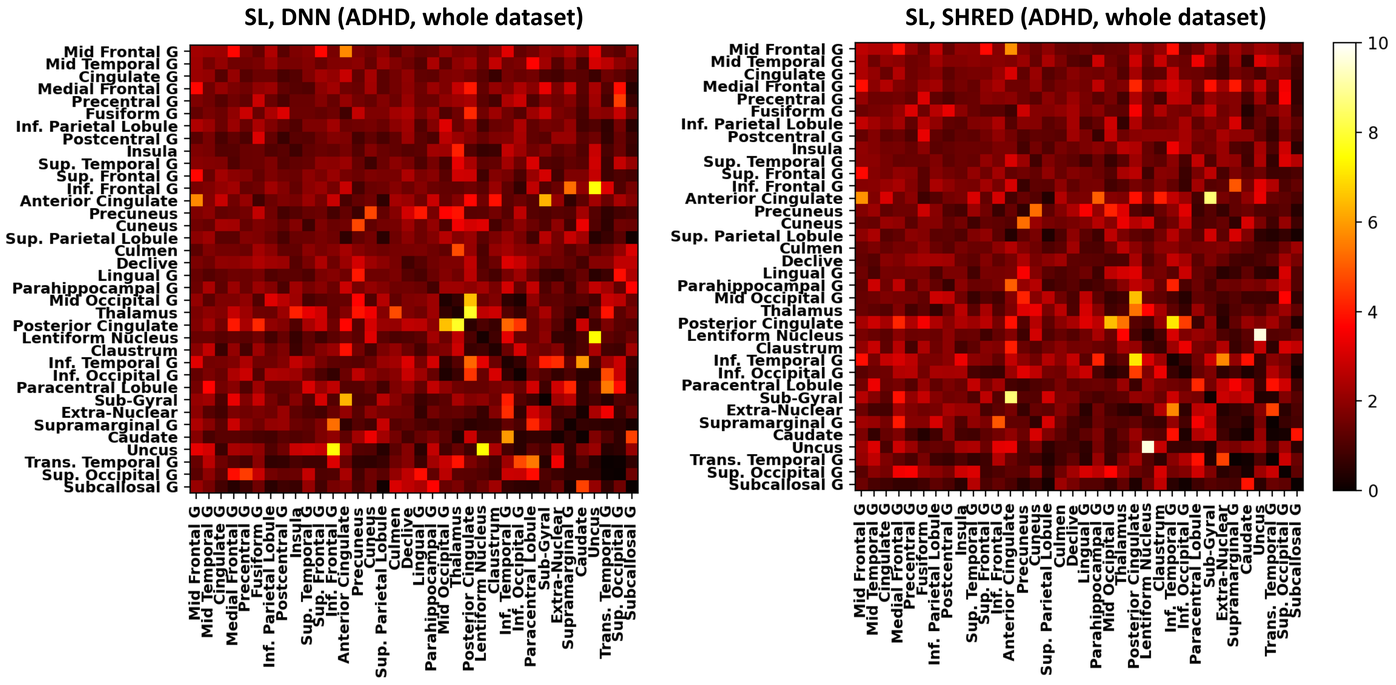
Done via **Integrated Gradients**, with normal controls as baseline.

$$R_v(x_j|F) = (x_{jv} - \bar{x})\int_0^1 \frac{\partial F\,\left(\bar{x} - \alpha(x_j - \bar{x})\right)}{\partial x_{jv}}\,d\alpha$$

$$R_{v_F} = \frac{1}{M}\sum_{j\in\{j|y_j=1\}}\left|R_v(x_j|F)\right|$$

Results

ABIDE	SL			SSL	SSL + Harmonisation	
	ASD-SAENet	GAE-FCNN	DNN	EDC-VAE	SHRED	SHRED-II
CALT	77.6 ± 4.5	65.2 ± 7.4	86.0 ± 2.9	88.6 ± 1.9	89.8 ± 2.6	**93.9 ± 1.3**
LEU1	76.5 ± 3.1	79.0 ± 2.5	80.3 ± 2.9	80.4 ± 3.9	**84.7 ± 3.7**	84.5 ± 3.0
LEU2	73.4 ± 2.8	73.0 ± 3.5	81.4 ± 2.4	86.7 ± 3.6	83.29 ± 2.7	**95.7 ± 2.3**
MAXM	74.1 ± 3.5	70.1 ± 4.2	82.8 ± 3.0	84.3 ± 3.7	**87.6 ± 2.8**	86.1 ± 3.4
NYU	70.9 ± 2.2	71.6 ± 2.0	78.3 ± 1.3	79.1 ± 1.0	77.27 ± 1.5	**79.3 ± 1.2**
OHSU	74.3 ± 3.6	77.4 ± 6.5	90.5 ± 4.8	93.0 ± 3.1	**96.4 ± 3.6**	95.4 ± 3.8
OLIN	76.8 ± 6.2	85.2 ± 3.8	88.8 ± 3.2	90.8 ± 4.6	89.2 ± 4.2	**96.0 ± 2.7**
PITT	77.9 ± 4.3	71.8 ± 4.6	84.8 ± 2.2	87.7 ± 3.0	88.4 ± 3.3	**88.5 ± 1.9**
STAN	74.5 ± 4.5	73.4 ± 3.2	84.0 ± 4.9	86.5 ± 4.2	88.4 ± 4.1	**91.0 ± 3.8**
TRIN	74.4 ± 5.5	62.8 ± 2.5	80.6 ± 2.7	83.2 ± 1.8	85.0 ± 2.5	**88.7 ± 2.5**
UCLA1	74.4 ± 2.6	72.6 ± 3.5	78.7 ± 3.6	79.8 ± 3.1	79.3 ± 4.0	**81.3 ± 3.7**
UCLA2	67.5 ± 3.5	76.5 ± 4.7	84.0 ± 7.4	88.5 ± 6.3	86.0 ± 5.2	**90.5 ± 3.7**
UM1	71.3 ± 3.1	70.7 ± 2.9	78.2 ± 3.0	81.1 ± 3.0	80.0 ± 2.8	**82.9 ± 2.7**
UM2	67.2 ± 3.0	68.5 ± 2.7	85.1 ± 3.8	87.7 ± 3.4	87.6 ± 4.3	**94.1 ± 2.7**
USM	81.3 ± 3.1	73.7 ± 2.7	89.2 ± 1.7	90.6 ± 1.6	88.2 ± 2.0	88.1 ± 2.0
YALE	71.4 ± 3.5	65.7 ± 3.4	80.5 ± 3.4	84.8 ± 2.7	83.3 ± 3.5	**88.6 ± 3.0**
ADHD						
NI	80.0 ± 3.3	60.2 ± 2.3	79.1 ± 3.0	80.2 ± 4.9	83.1 ± 4.8	**83.1 ± 4.1**
NYU	63.9 ± 0.9	63.1 ± 1.9	63.6 ± 2.1	64.4 ± 2.1	**68.7 ± 1.6**	64.4 ± 2.2
OHSU	68.4 ± 1.8	67.3 ± 2.3	70.4 ± 3.1	71.4 ± 3.0	**75.0 ± 2.9**	72.6 ± 2.8
PKU	71.4 ± 1.5	67.8 ± 1.0	74.4 ± 1.6	75.3 ± 1.8	76.4 ± 1.0	**76.5 ± 1.5**



Quantitative evaluation: both SSL and data harmonisation led to improved performance.

Qualitative evaluation: DNN produces site-specific biomarkers, while SHRED produces site-invariant biomarkers present in all 4 ADHD sites.