POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

A Multi-center Normative Deep Learning Approach for Automatic Detection of Bipolar Disorder Patients based on Neuroanatomy

TESI MAGISTRALE IN BIOMECIAL ENGINEERING – INGEGNERIA BIOMEDICA

AUTHOR: INÊS WON SAMPAIO

ADVISOR: PROF. ELEONORA MAGGIONI

ACADEMIC YEAR: 2021-2022

1. Introduction

Bipolar Disorder (BD) is a chronic and disabling mood disorder with a lifetime prevalence of around 2-5%. People suffering from BD switch between depressive, maniac, and euthymia phases. However, the majority of BD patients suffer from depression at their first-lifetime affective episode and are initially misdiagnosed. Furthermore, BD heterogeneity has prevented the identification of specific neurobiological markers that could lead to an early, objective, and precise diagnosis of the disease. Structural magnetic resonance imaging (sMRI) data has been widely used to detect differences in white matter (WM) and grey matter (GM) morphology between HC and BD, but findings are fragmented.

Increased interest in Machine Learning (ML) has bloomed in the psychiatric disorder research field due to usefulness in assisting psychiatrists with diagnosis and prognosis. Nevertheless, reported results on the accuracy of BD diagnosis through ML analysis have been rather polarizing, ranging from 54.8% to 100% perhaps due to a lack of methodological standards, including data processing methods. Besides, many criticisms have been raising up, concerning the lack of domain relevance of most "black box" models, providing no insight related to the pathophysiology mechanisms. Recently, innovative approaches based on Deep Learning (DL) models using anomaly detection methods have successfully attenuated this issue [1]. Up to now, autoencoder (AE) normative models have been successfully used to detect alterations in brain morphology in autism and schizophrenia, showing their remarkable potential in the search for biomarkers of psychiatric disorders.

Within this context, this thesis aims to provide knowledge of neuroanatomical bases of BD in order to accurately and automatically recognize BD from healthy controls (HC). We use a multisite dataset composed of HC and BD samples, totaling 1163 subjects, from which we extract cortical thickness and volumetric regions of interest (ROI) features. An AE-based normative model is then developed in order to assess the possibility to automatically detect BD patients as *deviating* samples.

To design a reliable DL model, yielding clinical applicability, we addressed the following issues. Confounding effects like age, sex, total intracranial volume (TIV) and non-biological site effects, e.g., those associated with center-specific sMRI parameters, must be removed from data. Finally, an appropriate Cross-Validation (CV) framework

for internal validation and an external validation pipeline must be employed to achieve a robust models' generalization error estimation.

2. Aim of the work

We propose to investigate a normative approach for BD discrimination, by employing an AE model trained on HC data. A secondary objective of this work is to find a proper data processing pipeline, for which we evaluate different harmonization methods combined with biological covariates correction.
We use the trained model to reconstruct test set data and the reconstruction errors between HC and BD to evaluate models' discriminative performance. Then, we extract patterns of neuroanatomical deviations identified in the BD group and evaluate if this subset of features is generalizable and improves discriminative performance in an external independent set. We parallelly train an SVM model to classify BD to be use as baseline comparison, and we consider the ENIGMA study [2] results as most updated state-of-art for BD classification comparison.
Thus, the aims of this work are: to produce a successful normative model to reconstruct healthy brain features; to discriminate BD against HC using the normative model; to extract brain-feature abnormalities characterizing patients within the heterogeneous BD spectrum; to assess if BD can be classified by using the subset of unique relevant brain features instead of all brain features; assess any improvement in BD classification obtained using the normative-based approach with respect to the classical SVM classifier; identify the optimal site-effect removal pipeline to be integrated into a ML analysis.

3. Methods

MRI scans and pre-processing pipelines were performed in Matlab R2018a (The Mathworks, Inc®) environment. Data processing and ML pipelines, were built using Google Colaboratory with Python 3.7.13.

3.1. Dataset Description

The dataset used in this thesis work is composed of 605 HC subjects and 558 BD patients, gathered from 7 centers, whose description is reported in Table 3.1.

ID	Center	HC	BD	Total
1	AUOV	93	20	113
2	FSL_ ROME	250	257	507
3	JUH	111	23	134
4	MI_POLI	26	12	38
5	OSR	67	133	200
6	PITTS	28	58	86
7	UBC	30	55	85

Table 3.1 Dataset Description.

As reported in Table 3.2, the BD group is on average older than the HC group and within each group, there are slightly more females than male subjects.

		Training Set	Test Set	External Set
HC	Sex	230|289	27|31	13|15
	age	37.1±15.0	33.9±14.0	28.6±4.6
BD	Sex	-	34|41	24|34
	age	-	40.3±12.7	33.8±10.4

*Age in years; Sex numerosity reported as: Males | Females

Table 3.2 Dataset Demographic Description.

3.2. MRI pre-processing

The sMRI scans were acquired in the 7 centers using T1-weighted sequences on 3T RMN scanners. The raw MRI scans were processed using a gold-standard protocol. The Voxel-Based Morphometry (VBM) pre-processing was performed using the SPM12 Computational Anatomy Toolbox (CAT12) toolbox. The following GM morphological features were extracted through anatomical automatic labeling: 68 cortical thickness (CT) values from Desikan-Killiany atlas cortical regions and 52 GM volume (GMV) values from CoBra atlas subcortical regions.

3.3. Cross-Validation Framework

A center dataset was randomly holdout as an external set, PITTS center data, for an external validation of the neuroanatomical deviating features. For internal validation, the remaining 6 centers were split with an holdout method, stratifying for center proportions, where the training set is only composed of HC samples. The

HC dataset was split into 90% training (519) and 10% test (58) sets, and the BD test set was composed of 15% (75) randomly selected from the total BD dataset (500). The analysis of neuroanatomical deviating features, the feature selection step, was performed using the BD dataset composed of 500 subjects. A 10-fold CV was used for model optimization, for which only the training set was used, retrieved from the splitting previously described. The training set 10-fold CV splits were stratified for center proportions.
The best hyperparameter combination was chosen based on the lower mean reconstruction error, MSE, in the fold used for validation. Then, the model was retrained with the entire training set.

3.4. Modeling Confounding Variables

To correct for confounding signals encoded in the neuroimaging data, two separate regression methods were applied to control for non-biological site effects and biological covariates.

- **Data Harmonization**

The harmonization step should remove from data systematic non-biological differences from data that make samples not directly comparable due to the inter-site variability (i.e., batch effect) while preserving the association between data and biological covariates of interest. Thus, site effects were removed employing ComBat (Combatting Batch Effects) tool, an empirical Bayes framework [3]. Because there is no standardized approach for multisite data harmonization in a ML analysis, we specifically designed a pipeline that can be integrated into both internal and external ML validation frameworks, i.e., to only estimate effects in a training set and to apply them separately to a test set or external set. For the internal validation, we use the neurocombat function [4], provided in https://github.com/Jfortin1/ComBatHarmonization, available in the form *neuroCombatFromTraining*, for separate test set harmonization, which we have called the CV-ComBat option. For the external validation framework, we design a pipeline based on C. Stein et al.[5], M-ComBat function, which proposes to center data on a location and scale of a pre-determined batch reference. Thus we harmonize a posteriori an external set with neurocombat function, by setting the *reference_batch* option as the whole harmonized training set. We have named this approach as Ref-ComBat option.
CT and GMV features are harmonized separately. For CT features, age and sex are considered as the biological covariates, whereas for GMV features, we also include TIV. For the latter step, TIV is first itself harmonized with the GMV. Afterward, the original GMV features are harmonized considering age, sex, and harmonized TIV as biological covariates.

- **Regressing-out bio-covariates**

For the biological covariates removal we followed the CV method recommended in [6]. Linear regression is fitted to training data, considering each brain feature as dependent variable Y, and biological covariates as independent variables. We assume that age-related changes and inbetween sex differences are comparable between HC and BD. Data is standardized by estimating statistics in the training set, before the regression fit, and after removing the confounder effects.

- **Processing Pipelines**

We investigate four harmonization options: No harmonization (A), harmonizing within an internal validation framework, using CV-ComBat (B) and external validation framework using Ref-ComBat (C), and harmonizing the whole data set prior to dataset splitting (D). The 5 parallel processing pipelines which will be compared are then:
1) No Data Correction (A) Pipeline
Including correction for bio-covariates:
2) No Harmonization (A) Pipeline
3) Whole Dataset Harmonization (D) Pipeline
4) Whole Dataset Harmonization (D) + External Set Harmonization (C) Pipeline
5) CV-Harmonization (B+ C) Pipeline

3.5. AE Normative Model

- **AE-based model**

The AE is composed of 5 layers, including input and output with 120 hidden units. Table 3.3 shows the chosen fixed hyperparameters. Besides, the output layer is composed of a linear activation function using a Glorot uniform parameter initializer. The hyperparameters which were tuned were: layer dimensions 2,3,4 – constraining dim 2>dim3<dim4, L2 norm regularization technique (the same for all layers), and learning rate. The training process was allowed to stop when overfitting after more than 250 epochs, restoring the best model parameters.

AF	Loss	Optimizer	Batch	Epochs
SELU	MSE	Adam	35	2000

Table 3.3: Fixed Hyperparameters.

- **Model Evaluation**

All processing pipelines were tested with the following framework. The test set is passed through the trained model and a reconstruction error score is attributed to each subject, a Deviation Metric (DM), equal to the reconstruction MSE, averaging all features per subject. The group comparisons are performed by employing a one-sided Mann-Whitney U (MWU) test applied to the DM, assuming the alternative hypothesis of BD-DM to be greater than HC-DM ($p<0.05$). A ROC curve is carried out using the subject DM data and the diagnosis as the binary target variable, 1 for BD and 0 for HC.

- **Feature Selection**

In this step, we use all BD subjects' dataset (n=500 samples). Each brain regional feature is considered alone and its square reconstruction error calculated for each BD subject and for each HC subject. Then, the two groups are compared for each feature with a one-sided MWU-Test. The brain regional features that are found to be associated with a significant p-value (i.e. $p< 0.05$) indicate that their reconstruction error was significantly greater in the BD group. The AUC-ROC curve is re-performed considering only this subset of features, a form of circular analysis, but confirms whether they improve the models' discriminative power.

- **Classification**

The BD classification is performed considering the previous feature subset. The features are selected in the test set and thus must be validated in an external independent set. The PITTS external set is passed through the network and the subject DM is calculated, by averaging only the subset of features. If the feature subset is generalizable, the discriminative performance of the model using these features should be comparable between the test set and the external set.

3.6. SVM model

From the 7 centers contained in the dataset, one out of four is holdout as an external site set, specifically all data from MI_POLI, OSR, PITTS, and UNC sites, thus following a LOSO-CV framework, for which processing pipeline 1 (i.e., no data correction, A) and 5 (CV-Harmonization, B+C) are tested. Afterward, for each LOSO trial, the rest of the dataset, 6 centers, was split into a 70% training set and a 30% test set, stratifying for center proportions. Since data on the training set included both HC and BD, the diagnosis is included as a biological covariate in the harmonization with ComBat. The SVM model used was the one reported in the ENIGMA Study [2] which uses a linear kernel and parameter C=1.

4. Results

4.1. Model Optimization

From the hyperparameter combinations reported in Table 4.1. The combination yielding the lower reconstruction error in the training set folds was: Layer 2,4=100, Layer 3=85, L2_regul=1e-4, Lr=1e-4. The learning rate schedule, denoted as *lr_schedule* in the table above had initial learning rate= 0.001 and decay step= 0.9977 .

Layer 2,4	Layer 3	L2 regularizer	Learning rate
100	80,75, 60	1_{10}-5,1_{10}-4, 1_{10}-3, 0.01	1_{10}-4, 1_{10}-3, 1_{10}-2, lr_schedule
100	85,70, 65	1_{10}-5, 1_{10}-4	1_{10}-4, lr_schedule
80	75	1_{10}-5,1_{10}-4, 1_{10}-3, 0.01	1_{10}-4, 1_{10}-3, 1_{10}-2, lr_schedule

Table 4.1 Hyperparameter Grid.

4.2. AE Normative Approach

The results will be reported in detail for pipeline 5, CV Harmonization (option B +C), because we consider and suggest that the processing pipeline integrated into the CV framework is the most rigorous one. A summary table, reported in Table 4.2, shows the results for all pipelines.

The best model gave an average reconstruction error on the training set (i.e., only HC) of 0.0278 and 0.0710 in the HC test set. The MWU test performed on BD and HC DM gave p-value=0.282 (statistic=15172), showing BD DM was not significantly greater than HC DM. The AUC-ROC curve result, using the BD test set DM (15% of all BD subjects) and HC test set DM was 0.51, in the chance line. The models' reconstruction error didn't yield any discriminative performance in the test set.

In the feature selection procedure, the regions found deviating significantly in the BD group

were: left medial orbital frontal, left superior parietal, left superior posterior cerebellar lobule VI, left hippocampus CA1, right globus pallidus, and right amygdale. The subject DM was again calculated in the test set, considering this subset of features, and an MWU test was performed resulting in a p-value=.001 and an AUC-ROC curve improving to 0.66.
To assess the generalizability of the subset of features they are tested using the PITTS external set, which is passed through the model, and respective subject DMs are calculated, for both all features and in the subset of features, resulting in an AUC=0.58 and AUC=0.61 respectively.

Pipeline	Test Set		External Set	
	All feat. AUC	Subset feat. AUC	All feat. AUC	Subset feat. AUC
1	0.56	0.69	0.45	0.51
2	0.56	0.72	0.39	0.43
3	0.52	0.61	0.91	0.71
4	0.50	0.63	0.45	0.54
5	**0.51**	**0.66**	**0.58**	**0.61**

Table 4.2: Processing Pipelines Results.

4.3. SVM model

The average results for the LOSO-CV analysis are reported in Table 4.3 in the two first rows, while the third represents the unique result for pipeline 5 when the PITTS set is considered the external set.
For the site-level analysis, the AUC-ROC curve ranged from 0.25 to 0.91.

Pipeline	Test set AUC	External Set AUC
1	0.6300±0.0158	0.5050±0.0918
5	0.5350± 0.0150	0.5125±0.0621
5(PITTS)	0.55	0.50

Table 4.3: LOSO-CV Results.

5. Discussion

Regarding the AE-based normative approach, we can verify in pipeline 1, and pipeline 2, results that the model fails to generalize to the external set. The AUC results drop to the chance line or below-chance line from the test set (AUC_1=0.56, AUC_2=0.56), to the external set (AUC_1=0.45 and AUC_2= 0.39). Interestingly, comparing pipelines 1 and 2 external set AUCs the performance worsened in the second. Seems that removing biological covariates effects from non-harmonized data resulted in better performance in the test set but worst in the external set PITTS, which could be explained by a covariate shift in the latter set. In pipeline 3, WD- harmonization (D), the external set is harmonized outside the external validation framework, which breaks its independence from training and test set, while in pipeline 4 (option D+C), the external set is kept separated and harmonized a posteriori with ref-ComBat option. We verify that the good performance in the external set in pipeline 3 drops to chance-line in pipeline 4 (AUC_3=0.92, AUC_4=0.45), which shows how indirect data leakage could positively bias the results. Finally, in pipeline 5, CV-Harmonization (option B+C), we verify a good generalization performance to the external set (AUC_test=0.51, AUC_ext=0.58), although the result in the test set depicts a lack of discriminative capability of the models' reconstruction error. Besides, the latter improved AUC in the external set might be explained by some remaining heterogeneity in the test set, composed of data from 6 centers that were randomly split, which due to an average effect cancels out above-chance performances, while the external set data is fully homogeneous. Moreover, assessing the results of the feature subset generalization we verify that the circular analysis in the test set gave an AUC=0.66, improved from AUC=0.51, and the test in the external set resulted in an AUC=0.61, improved from AUC=0.58 using all features thus, showing that the feature subset was generalized to the external set. Comparing the normative approach performance for all features with our SVM model, for pipeline 5, we can conclude that in the test set they show comparable AUCs results, on chance-line, while in the PITTS external set the normative approach yields better generalization performance (AUC_SVM=0.50, AUC_AE=0.58). Nevertheless, we verify that our SVM model had worst results than the SVM model in the ENIGMA study, which reported a LOSO-CV AUC=60.92 compared to our AUC=53.00, and a site-level analysis with an AUC ranging from 40.00 to 71.00, while we achieved AUC ranging from 25.00 to 91.00 [2]. The key methodological differences that might have hampered the performance of the SVM model in our analysis concern mainly the data numerosity.

Finally, performing a recall-precision curve with the external set subject's DM for the feature subset, in pipeline 5, we extract the optimal result, yielding a recall of 0.40 and a precision of 0.85. Since the average psychiatrist's sensitivity (also called recall) in diagnosing BD is estimated at 31%, we argue that the normative approach yields promising results.

6. Conclusions

We had proposed to investigate different processing pipelines and normative approach performance for BD discrimination. Regarding the first, we can conclude that the CV-Harmonization (option B+C), used in pipeline 5, was effective in harmonizing data and is the most rigorous option to use, as it can be integrated into both internal and external validation frameworks. Using the latter option, we verify the best generalization results to the external set. We conclude that harmonizing data helps improve models' generalization capability while not doing so leads the model to have good performances in an internal validation framework (test set) but failing to generalize to an external set (pipelines 1 and 2). We also show the dangers of performing data processing steps outside the validation frameworks in an ML analysis, by the performance in pipeline 3 and consequent drop in pipeline 4. Finally, we can conclude that our proposed methodological approach for BD classification yields promising results, as the neuroanatomical deviating features selected in the test set were generalizable to the external set. Nevertheless, future development of this work would be to perform a LOSO-CV and a nested 10-fold CV for the AE-based model, to have results yielding higher statistical power in the normative approach. We also conclude that the model was not able to discriminate BD when considering all features. The latter could be a consequence of the heterogeneity of BD, deeming the anomaly detection approach inappropriate to detect BD. However, it could also be a consequence of a poor optimization strategy, since minimizing reconstruction error does not necessarily lead to maximizing anomaly detection performance. In fact, our normative model yields good reconstruction performances for both HC and BD. Future development would be to explore improved optimization strategies to achieve better discrimination based on the reconstruction error.

References

[1] W. H. L. Pinaya, A. Mechelli, and J. R. Sato, "Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study," *Human Brain Mapping*, vol. 40, no. 3. pp. 944–954, 2019, doi: 10.1002/hbm.24423.

[2] A. Nunes *et al.*, "Using structural MRI to identify bipolar disorders – 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group," *Mol. Psychiatry*, vol. 25, no. 9, pp. 2130–2143, 2020, doi: 10.1038/s41380-018-0228-9.

[3] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007, doi: 10.1093/biostatistics/kxj037.

[4] J. P. Fortin *et al.*, "Harmonization of cortical thickness measurements across scanners and sites," *Neuroimage*, vol. 167, no. June 2017, pp. 104–120, 2018, doi: 10.1016/j.neuroimage.2017.11.024.

[5] C. K. Stein *et al.*, "Removing batch effects from purified plasma cell gene expression microarrays with modified ComBat," *BMC Bioinformatics*, vol. 16, no. 1, pp. 1–9, 2015, doi: 10.1186/s12859-015-0478-3.

[6] L. Snoek, S. Miletić, and H. S. Scholte, "How to control for confounds in decoding analyses of neuroimaging data," *Neuroimage*, vol. 184, no. September 2018, pp. 741–760, 2019, doi: 10.1016/j.neuroimage.2018.09.074.

7. Acknowledgements

I would like to thank my adviser Prof. Eleonora Maggioni, for all the support and constructive feedback, to my co-adviser Emma Tassi for helping me and follow attentively my work and to my co-adviser Prof. Paolo Brambilla, without whom this thesis project opportunity would have not been possible.