Published in final edited form as:

Mach Learn Med Imaging. 2017 September ; 10541: 371–378. doi:10.1007/978-3-319-67389-9_43.

Machine Learning for Large-Scale Quality Control of 3D Shape Models in Neuroimaging

A full list of authors and affiliations appears at the end of the article.

Abstract

As very large studies of complex neuroimaging phenotypes become more common, human quality assessment of MRI-derived data remains one of the last major bottlenecks. Few attempts have so far been made to address this issue with machine learning. In this work, we optimize predictive models of quality for meshes representing deep brain structure shapes. We use standard vertex-wise and global shape features computed homologously across 19 cohorts and over 7500 human-rated subjects, training kernelized Support Vector Machine and Gradient Boosted Decision Trees classifiers to detect meshes of failing quality. Our models generalize across datasets and diseases, reducing human workload by 30–70%, or equivalently hundreds of human rater hours for datasets of comparable size, with recall rates approaching inter-rater reliability.

Keywords

shape analysis; machine learning; quality control

1 Introduction

In recent years, large-scale neuroimaging studies numbering in the thousands and even 10's of thousands of subjects have become a reality [1]. Though automated MRI processing tools [2] have become sufficiently mature to handle large datasets, visual quality control (QC) is still required. For simple summary measures of brain MRI, QC may be a relatively quick process. For more complex measures, as in large studies of voxel- and vertex-wise features [3], the QC process becomes more time-intensive for the human raters. Both training of raters and conducting QC ratings, once trained, can take hours even for modest datasets.

This issue is particularly relevant in the context of multi-site meta-analyses, exemplified by the ENIGMA consortium [1]. Such studies, involving dozens of institutions, require multiple researchers to perform quality control on their cohorts, as individual data cannot always be shared. In addition, for meta-analysis studies performed after data collection, the QC protocols must be reliable in spite of differences in scanning parameters, post-processing, and demographics. In effect, QC has become one of the main practical bottlenecks in big-data neuroimaging. Reducing human rater time via predictive modeling and automated quality control is bound to play an increasingly important role in maintaining and hastening the pace of the scientific discovery cycle in this field.

*these authors contributed equally

In this paper, we train several predictive models for deep brain structure shape model quality. Our data is comprised of the ENIGMA Schizophrenia and Major Depressive Disorder working groups participating in the ENIGMA-Shape project [3]. Using ENIGMA's Shape protocol and rater-labeled shapes, we train a discriminative model to separate "FAIL"(F) and "PASS"(P) cases. For classification, we use a support vector classifier with a radial basis kernel (SVC) and Gradient Boosted Decision Trees (GBDT). Features are derived from the standard vertex-wise measures as well as global features. For six out of seven deep brain structures, we are able to reduce human rater time by 30 to 70 percent in out-of-sample validation, while maintaining FAIL recall rates similar to human inter-rater reliability. Our models generalize across datasets and disease samples.

2 Methods

Our goal in using machine learning for automated QC differs somewhat from most predictive modeling problems. Typical two-class discriminative solutions seek to balance misclassification rates of each class. In the case of QC, we focus primarily on correctly identifying FAIL cases, by far the smaller of the two classes (Table 1). In this first effort to automate shape QC, we do not attempt to eliminate human involvement, but simply to reduce it by focusing human rater time on a smaller subsample of the data containing nearly all failing cases. Our quality measures, described below, reflect this nuance.

2.1 MRI processing and shape features

Our deep brain structure shape measures are computed using a previously described pipeline [4,5], available via the ENIGMA Shape package. Briefly, structural MR images are parcellated into cortical and subcortical regions using FreeSurfer. Among the 19 cohorts participating in this study, FreeSurfer versions 5.1 and 5.3 were used, depending on the institution. The binary region of interest (ROI) images are then surfaced with triangle meshes and parametrically (spherically) registered to a common region-specific surface template [6]. This leads to a one-to-one surface correspondence across the dataset at roughly 2,500 vertices per ROI. Our ROIs include the left and right thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and nucleus accumbens. Each vertex p of mesh model M is endowed with two shape descriptors:

Medial Thickness, $D(p) = \|c_p - p\|$, where c_p is the point on the medial curve c closest to p.

$LogJac(p)$, Log of the Jacobian determinant J arising from the template mapping, $J : T_{\phi(p)}M_t \rightarrow T_pM$.

Since the ENIGMA surface atlas is in symmetric correspondence, i.e. the left and right shapes are vertex-wise symmetrically registered, we can combine two hemispheres for each region for the purposes of predictive modeling. At the cost of assuming no hemispheric bias in QC failure, we effectively double our sample.

The vertex-wise features above are augmented with their volume-normalized counterparts: $\{D, J\}_{normed}(p) = \frac{\{D,J\}(p)}{V^{\{\frac{1}{3},\frac{2}{3}\}}}$. Given discrete area elements of the template at vertex p, $A_t(p)$,

we estimate volume as $V = \sum_{p \in vrts(M)} 3A_t(p)J(p)D(p)$. We also use two global features: the shape-wide feature median, and the shapewise 95th percentile feature threshold.

2.2 Human quality rating

Human-rated quality control of shape models is performed following the ENIGMA-Shape QC protocol[44]. Briefly, raters are provided with several snapshots of each region model as well as its placement in several anatomical MR slices (Fig. 1). A guide with examples of FAIL (QC=1) and PASS (QC=3) cases is provided to raters, with an additional category of MODERATE PASS (QC=2) suggested for inexperienced raters. Cases from the last category are usually referred to more experienced raters for second opinions. Once a rater becomes sufficiently experienced, he or she typically switches to the binary FAIL/PASS rating. In this work, all remaining QC=2 cases are treated as PASS cases, consistent with ENIGMA shape studies.

2.3 Predictive models

First, we used Gradient Boosted Decision Trees (GBDT). This is a powerful ensemble learning method introduced by Friedman [7] in which subsequent trees correct for the errors of the previous trees. In our experiments we used the Xgboost [8] implementation due to speed and regularization heuristics, with the logistic loss function. Second, we used Support Vector Classifier. Based on earlier experiments and the clustered nature of FAIL cases in our feature space, we used the radial basis function (RBF) kernel in our SVC models. Indeed, in preliminary experiments RBF outperformed linear and polynomial kernels. We used scikit-learn's [9] implementation of SVC.

2.4 Quality measures

In describing our quality measures below, we use the following definitions. TF stands for TRUE FAIL, FF stands for FALSE FAIL, TP stands for TRUE PASS, and FP stands for FALSE PASS. Our first measure, $\mathbf{F-recall} = \frac{TF}{TF + FP}$, shows the proportion of FAILS that are correctly labeled by the predictive model. The second measure, $\mathbf{F-share} = \frac{TF + FF}{\text{Number of observations}}$, shows the proportion of the test sample labeled as FAIL by the model. Finally, we used a **modified F-score**, which allows us to compare models based on the specific requirements of our task, i.e. a very high F-recall and F-share substantially below 1, we use a variation on the standard F-score.

$$\text{F-score}_{\text{mod}} = 2 \times \frac{\text{F-recall} \times (1 - \text{F-share})}{\text{F-recall} \times (1 - \text{F-share})}.$$

Note that the modified F-score cannot equal 1, as in the standard case. An ideal prediction leads to $\text{F-score}_{\text{mod}}$ = 1 - F-share. The intuition behind our custom F-score is based on the highly imbalanced FAIL and PASS samples. A model that accurately labels all failed cases

[44] enigma.usc.edu/ongoing/enigma-shape-analysis

is only valuable if it substantially reduces the workload for human raters, a benefit reflected by F-share.

3 Experiments

For each of the seven ROIs, we performed eight experiments defined by two predictive models (SVC and GBDT), two types of features (original and normed) and two cross-validation approaches. We tested ”Leave-One-Site-Out” and 5-fold stratified cross-validation, as described below.

3.1 Datasets

In our experiments, we used deep brain structure shape data from the ENIGMA Schizophrenia and Major Depressive Disorder working groups.

Our predictive models were trained using 15 cohorts totaling 5718 subjects’ subcortical shape models from the ENIGMA-Schizophrenia working group. The ENIGMA-Schizophrenia (ENIGMA-SCZ) working group is comprised of over two dozen cohorts from around the world. The goal of the working group is to identify subtle effects of Schizophrenia and related clinical factors on brain imaging features. For a complete overview of ENIGMA-SCZ projects and cohort details, see [10].

To test our final models, we used data from 4 cohorts in the Major Depressive disorder working group (ENIGMA-MDD), totaling 1509 subjects, for final out-of-fold testing. A detailed description of the ENIGMA-MDD sites and clinical questions can be found here [11].

3.2 Model validation

All experiments were performed separately for each ROI. The training dataset was split into two halves referred to as ’TRAIN GRID’ and ’TRAIN EVAL.’ The two halves contained data from each ENIGMA-SCZ cohort, stratified by the cohort-specific portion of FAIL cases. Model parameters were optimized using a grid search within ’TRAIN GRID’, with either stratified 5-fold or Leave-One-Site-Out cross-validation. Parameters yielding the highest Area Under the ROC-curve were selected from among all cross-validation and feature types.

Both SVC and GBDT produce probability estimates indicating the likelihood that the individual subject’s ROI mesh is a FAIL case, P_{FAIL}. Exploiting this, we sought a probability threshold for each model selected during the grid search to optimize the modified F-score in the ’TRAIN EVAL’ sample. This amounts to a small secondary grid search. To simplify traversing this parameter space, we instead sample F-score$_{mod}$ at regularly spaced values of P_{FAIL}, from 0.1 to 0.9 in 0.1 increments. This is equivalent to F-share in the ’TRAIN EVAL’ sample (Eval F-share, Table 2).

Final thresholds (Thres in Table 2) were selected based on the highest F-score$_{mod}$, requiring that F-recall ≥ 0.8 - a minimal estimate of inter-rater reliability. It is important to stress that

while we used sample distribution information in selecting a threshold, the final out-of-sample prediction is made on an individual basis for each mesh.

4 Results

Trained models were deliberately set to use a loose threshold for FAIL detection, predicting 0.3–0.8 of observations as FAILs in the TRAIN GRID sample. These predicted FAIL observations contained 0.85–0.9 of all true FAILs, promising to reduce the human rater QC time by 20–70%. These results largely generalized to the 'TRAIN EVAL' and test samples: Table 2 shows our final model and threshold performance for each ROI.

With the exception of the thalamus, our final models' performance measures generalized to the test sample, in some cases having better sample F-recall and lower percentage of images still requiring human rating compared to the evaluation sample. A closer look suggests that variability in model predictions across sites generally follows human rater differences. Table 3 breaks down performance by test cohort. It is noteworthy that the largest cohort, Münster (N = 1033 subjects, 2066 shape samples), has the best QC prediction performance.

At the same time, the "cleanest" dataset, Houston, with no human-detected quality failures, has the lowest F-share. In other words, Houston would require the least human rater time relative to its size, as would be hoped.

Visualizing the test results in Figure 2, we see the trend for lower F-share with higher overall dataset quality maintained by the smaller cohorts, but reversed by Münster. This could be a reflection of our current models' bias toward accuracy in lower-quality data due to greater numbers of FAIL examples (i.e., FAILs in high and low quality datasets may be qualitatively different). At the same time, F-recall appears to be independent of QC workload reduction due to ML, with most rates above the 0.8 mark.

5 Conclusion

We have presented a preliminary study of potential machine learning solutions for semi-automated quality control of deep brain structure shape data. Though some work on automated MRI QC exists [12], we believe this is the first ML approach in detecting end-of-the-pipeline feature failure in deep brain structure geometry. We showed that machine learning can robustly reduce human visual QC time for large-scale analyses for six out of the seven regions in question, across diverse MRI datasets and populations. Failure of the thalamus ML QC ratings to generalize out-of-sample may be explained by the region's specific features. Though we have only used geometry information in model training, MRI intensity, available to human raters for all ROI's, plays a particularly important role in thalamus ratings. The most common thalamus segmentation failure is the inclusion of lateral ventricle by FreeSurfer. Geometry is generally altered undetectably in such cases.

Beyond adding intensity-based features, possible areas of future improvement include combining ML algorithms, exploiting parametric mesh deep learning, employing geometric data augmentation, and refining the performance measures. Specifically, mesh-based convolutional neural nets can help visualize problem areas, which can be helpful for raters.

Very large-scale studies, such as the UK Biobank, ENIGMA, and others, are becoming more common. To make full use of these datasets, it is imperative to maximally automate the quality control process that has so far been almost entirely manual in neuroimaging. Our work here is a step in this direction.

Authors

Dmitry Petrov[*,1,2], Boris A. Gutman[*,1], Shih-Hua (Julie) Yu[1], Theo G.M. van Erp[3], Jessica A. Turner[5], Lianne Schmaal[23,24], Dick Veltman[24], Lei Wang[4], Kathryn Alpert[4], Dmitry Isaev[1], Artemis Zavaliangos-Petropulu[1], Christopher R.K. Ching[1], Vince Calhoun[39], David Glahn[6], Theodore D. Satterthwaite[7], Ole Andreas Andreasen[8], Stefan Borgwardt[9], Fleur Howells[10], Nynke Groenewold[10], Aristotle Voineskos[11], Joaquim Radua[12,33,34,35], Steven G. Potkin[3], Benedicto Crespo-Facorro[13,37], Diana Tordesillas-Gutiérrez[13,37], Li Shen[14], Irina Lebedeva[15], Gianfranco Spalletta[16], Gary Donohoe[17], Peter Kochunov[18], Pedro G.P. Rosa[19,32], Anthony James[20], Udo Dannlowski[25], Bernhard T. Baune[30], André Aleman[31], Ian H. Gotlib[26], Henrik Walter[27], Martin Walter[28,40,41], Jair C. Soares[29], Stefan Ehrlich[42], Ruben C. Gur[7], N. Trung Doan[8], Ingrid Agartz[8], Lars T. Westlye[8,36], Fabienne Harrisberger[9], Anita Riecher-Rössler[9], Anne Uhlmann[10], Dan J. Stein[10], Erin W. Dickie[11], Edith Pomarol-Clotet[12,33], Paola Fuentes-Claramonte[12,33], Erick Jorge Canales-Rodríguez[12,33,38], Raymond Salvador[12,33], Alexander J. Huang[3], Roberto Roiz-Santiañez[13,37], Shan Cong[14], Alexander Tomyshev[15], Fabrizio Piras[16], Daniela Vecchio[16], Nerisa Banaj[16], Valentina Ciullo[16], Elliot Hong[18], Geraldo Busatto[19,32], Marcus V. Zanetti[19,32], Mauricio H. Serpa[19,32], Simon Cervenka[21], Sinead Kelly[22], Dominik Grotegerd[25], Matthew D. Sacchet[26], Ilya M. Veer[27], Meng Li[28], Mon-Ju Wu[29], Benson Irungu[29], Esther Walton[42,43], and Paul M. Thompson[1] **for the ENIGMA consortium**

Affiliations

[1]Imaging Genetics Center, Stevens Institute for Neuroimaging and Informatics, University of Southern California, Marina Del Rey, CA, USA [2]The Institute for Information Transmission Problems, Moscow, Russia [3]Department of Psychiatry and Human Behavior, University of California Irvine, Irvine, CA, USA [4]Department of Psychiatry, Northwestern University, Chicago, IL, USA [5]Psychology Department & Neuroscience Institute, Georgia State University, Atlanta GA, USA [6]Yale University School of Medicine, New Haven, CT, USA [7]Department of Psychiatry, University of Pennsylvania School of Medicine, Philadelphia, PA, USA [8]CoE NORMENT, KG Jebsen Centre for Psychosis Research, Division of Mental Health and Addiction, Oslo University Hospital & Institute of Clinical Medicine, University of Oslo, Oslo, Norway [9]Department of Psychiatry, University of Basel, Basel, Switzerland [10]MRC Unit on Risk & Resilience to Mental Disorders, Department of Psychiatry and Mental Health, University of Cape Town, Cape Town, South Africa [11]Centre for Addiction and Mental Health, Toronto, Canada [12]FIDMAG Germanes Hospitalaries Research Foundation, Barcelona, Spain [13]University Hospital Marqués de Valdecilla, IDIVAL, Department of Psychiatry, School of Medicine, University of Cantabria, Santander, Spain [14]Department of Radiology and Imaging

Sciences, Indiana University School of Medicine, Indianapolis, IN, USA [15]Mental Health Research Center, Moscow, Russia [16]Laboratory of Neuropsychiatry, Santa Lucia Foundation IRCCS, Rome, Italy [17]School of Psychology, NUI Galway, Galway, Ireland [18]Maryland Psychiatric Research Center, University of Maryland School of Medicine, Baltimore [19]Department of Psychiatry, Faculty of Medicine, University of São Paulo, São Paulo, Brazil [20]University of Oxford, Oxford, United Kingdom [21]Centre for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden [22]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA [23]Orygen, The National Centre of Excellence in Youth Mental Health, Melbourne, Australia [24]Department of Psychiatry, VU University Medical Center, Amsterdam, The Netherlands [25]Department of Psychiatry and Psychotherapy, University of Münster, Germany [26]Department of Psychology, Stanford University, Stanford, CA, USA [27]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Department of Psychiatry and Psychotherapy, CCM, Berlin, German [28]Clinical Affective Neuroimaging Laboratory, Leibniz Institute for Neurobiology, Magdeburg, Germany [29]University of Texas Health Science Center at Houston, Houston, TX, USA [30]Discipline of Psychiatry, Adelaide Medical School, The University of Adelaide [31]Interdisciplinary Center Psychopathology and Emotion regulation (ICPE), Neuroimaging Center (BCN-NIC), University Medical Center Groningen, University of Groningen, Groningen, The Netherlands [32]Center for Interdisciplinary Research on Applied Neurosciences (NAPNA), University of São Paulo, São Paulo, Brazil [33]CIBERSAM, Centro Investigación Biomédica en Red de Salud Mental, Barcelona, Spain [34]Department of Clinical Neuroscience, Centre for Psychiatric Research, Karolinska Institutet, Stockholm, Sweden [35]Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, United Kingdom [36]Department of Psychology, University of Oslo, Oslo, Norway [37]CIBERSAM, Centro Investigación Biomédica en Red Salud Mental, Santander, Spain [38]Radiology department, University Hospital Center (CHUV), Lausanne, Switzerland [39]The Mind Research Network, Albuquerque, NM, USA [40]Leibniz Institute for Neurobiology, Magdeburg, Germany [41]Department of Psychiatry and Psychotherapy, University of Tübingen, Tübingen, Germany [42]Division of Psychological and Social Medicine and Developmental Neurosciences, Faculty of Medicine, TU Dresden, Germany [43]Psychology Department, Georgia State University, Atlanta, GA, USA

Acknowledgments

This work was funded in part by NIH BD2K grant U54 EB020403, Russian Science Foundation grant 17-11-01390 and other agencies worldwide.

References

1. Thompson PM, Andreassen OA, Arias-Vasquez A, Bearden CE, Boedhoe PS, Brouwer RM, Buckner RL, Buitelaar JK, Bulaeva KB, Cannon DM. ENIGMA and the individual: Predicting

factors that affect the brain in 35 countries worldwide. Neuroimage. 2015; 145(Pt B):389–408. [PubMed: 26658930]

2. Fischl B. Freesurfer. Neuroimage. 2012; 62(2):774–781. [PubMed: 22248573]
3. Gutman B, Ching C, Andreassen O, Schmaal L, Veltman D, Van Erp T, Turner J, Thompson PM, et al. Harmonized large-scale anatomical shape analysis: Mapping subcortical differences across the ENIGMA Bipolar, Schizophrenia, and Major Depression working groups. Biological Psychiatry. 2017; 81(10):S308.
4. Gutman BA, , Jahanshad N, , Ching CR, , Wang Y, , Kochunov PV, , Nichols TE, , Thompson PM. Medial demons registration localizes the degree of genetic influence over subcortical shape variability: An n= 1480 meta-analysis. Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on, IEEE; 14021406
5. Roshchupkin* GV, Gutman* BA, et al. Heritability of the shape of subcortical brain structures in the general population. Nature Communications. 2016; 7:13738.
6. Gutman BA, , Madsen SK, , Toga AW, , Thompson PM. Volume 8159 of Lecture Notes in Computer Science Springer International Publishing; 2013 24 A Family of Fast Spherical Registration Algorithms for Cortical Shapes; 246257
7. Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 2001; 29(5):1189–1232.
8. Chen T, , Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16; New York, NY, USA: ACM; 2016 785794
9. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.
10. van Erp TGM, Hibar DP, et al. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. Molecular psychiatry. 2015
11. Schmaal L, Hibar DP, et al. Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the enigma major depressive disorder working group. Mol Psychiatry. 2017; 22(6):900–909. [PubMed: 27137745]
12. Esteban O, Birman D, Schaer M, Koyejo OO, Poldrack RA, Gorgolewski KJ. MRIQC: Predicting quality in manual MRI assessment protocols using no-reference image quality measures. 2017 bioRxiv.

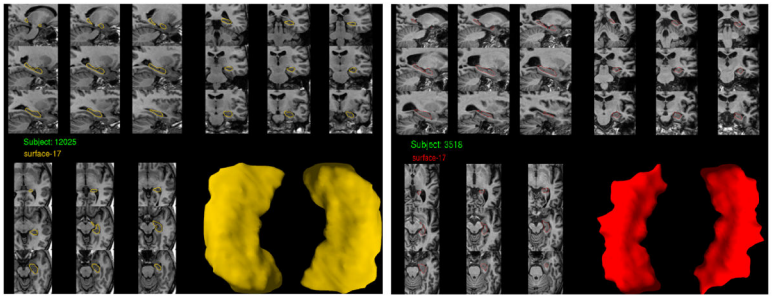

Fig. 1. Example hippocampal shape snapshots used for human QC rating
Left: A mesh passing visual QC. **Right:** A mesh failing visual QC.

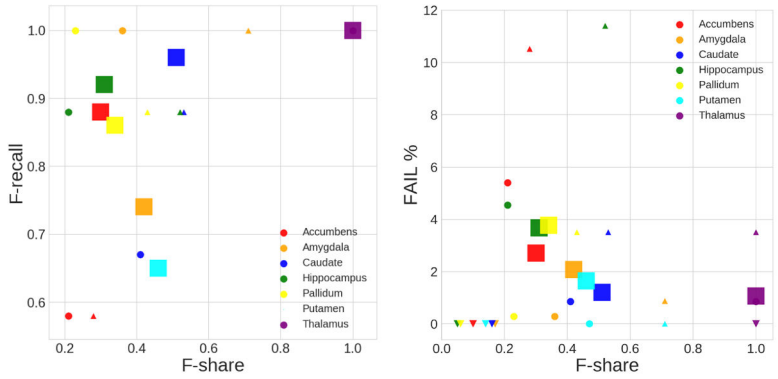

Fig. 2. Scatter plots of F-recall and actual FAIL case percentage vs. proportion of predicted FAIL cases on test datasets

Left: F-share vs F-recall. **Right:** Fail F-share vs FAIL percentage. Mark size shows the dataset size. Mark shape represents dataset (site): ○ **- CODE-Berlin (N=176);** □ **- Münster (N=1033);** △ **- Stanford (N=105);** ▽ **- Houston (N=195)**.

Table 1

Overview of FAIL percentage mean, standard deviation, maximum and minimum for each site. Sample sizes for each ROI vary slightly due to FreeSurfer segmentation failure.

	FAIL %	accumbens	caudate	hippocampus	thalamus	putamen	pallidum	amygdala
Train	mean±std	3.4±4.7	0.9±0.7	2.0±1.1	0.8±1.0	0.6±0.6	2.3±3.6	0.9±0.9
	max	16.4	2.1	4.2	3.4	1.5	13.8	2.6
	min	0.0	0.0	0.5	0.0	0.0	0.0	0.0
	size	10431	10433	10436	10436	10436	10435	10436
Test	mean±std	4.7±4.5	1.4±1.5	4.9±4.8	1.4±1.5	0.4±0.8	1.9±2.0	0.8±0.9
	max	10.5	3.5	11.4	3.5	1.6	3.8	2.1
	min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	size	3017	3018	3018	3018	3017	3018	3018

Table 2

Test performance of the models with the best F-score$_{mod}$ on evaluation (TRAIN EVAL). Excepting the thalamus, overall models' performance generalizes to out-of-sample test data

ROI	Model	CV	Features	Thres	Eval F-recall	Eval F-share	Eval F-score	Test F-recall	Test F-share	Test F-score
Accumbens	GBDT	5-fold	Normed	0.014	0.83	0.2	0.81	0.75	0.27	0.74
Amygdala	GBDT	5-fold	Normed	0.007	0.89	0.4	0.72	0.76	0.40	0.67
Caudate	SVC	LOSO	Original	0.017	0.84	0.3	0.76	0.92	0.45	0.68
Hippocampus	GBDT	5-fold	Normed	0.010	0.85	0.2	0.83	0.91	0.29	0.80
Pallidum	GBDT	5-fold	Normed	0.009	0.86	0.2	0.83	0.86	0.30	0.77
Putamen	SVC	LOSO	Normed	0.007	0.84	0.4	0.70	0.65	0.44	0.60
Thalamus	SVC	LOSO	Original	0.007	0.84	0.4	0.70	1.00	1.00	0.00

Table 3

Performance of best models for each test site. Models are the same as in Table 2. Symbol '-' indicates that there were no FAILs for particular ROI and test site.

ROI	Berlin F-recall	Berlin F-share	Stanford F-recall	Stanford F-share	Munster F-recall	Munster F-share	Houston F-recall	Houston F-share
Accumbens	0.58	0.21	0.58	0.28	0.88	0.30	-	0.10
Amygdala	1.00	0.36	1.00	0.71	0.74	0.42	-	0.17
Caudate	0.67	0.41	0.88	0.53	0.96	0.51	-	0.16
Hippocampus	0.88	0.21	0.88	0.52	0.92	0.31	-	0.05
Pallidum	1.00	0.23	0.88	0.43	0.86	0.34	-	0.06
Putamen	-	0.47	-	0.71	0.65	0.46	-	0.14
Thalamus	1.00	1.00	1.00	1.00	1.00	1.00	-	1.00