Interpretable Automatic Detection of Incomplete Hippocampal Inversions Using Anatomical Criteria

Lisa J. Hemforth[a], Claire Cury[b], Vincent Frouin[c], Sylvane Desrivières[e,f] , Antoine Grigis[c], Hugh Garavan[g], Rüdiger Brühl[h], Jean-Luc Martinot[i], Marie-Laure Paillère Martinot[i,j] , Eric Artiges[i,k] , Luise Poustka[m], Sarah Hohmann[d], Sabina Millenet[d], Nilakshi Vaidya[o], Henrik Walter[p], Robert Whelan[q], Gunter Schumann[p,r], Baptiste Couvy-Duchesne[a,s], Olivier Colliot[a], and The IMAGEN consortium[1]

[a]Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France
[b]Univ Rennes, CNRS, Inria, Inserm, IRISA UMR 6074, Empenn - ERL U 1228, F-35000 Rennes, France
[c]NeuroSpin, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France
[d]Department of Child and Adolescent Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Square J5, 68159 Mannheim, Germany
[e]Department of Neuroimaging, Institute of Psychiatry, Psychology & Neuroscience, King's College London, United Kingdom
[f]Centre for Population Neuroscience and Precision Medicine (PONS), Institute of Psychiatry, Psychology & Neuroscience, SGDP Centre, King's College London, United Kingdom
[g]Departments of Psychiatry and Psychology, University of Vermont, 05405 Burlington, Vermont, USA
[h]Physikalisch-Technische Bundesanstalt (PTB), Braunschweig and Berlin, Germany
[i]Institut National de la Santé et de la Recherche Médicale, INSERM U A10 "Trajectoires développementales en psychiatrie"; Université Paris-Saclay, Ecole Normale supérieure Paris-Saclay, CNRS, Centre Borelli; Gif-sur-Yvette, France
[j]AP-HP. Sorbonne Université, Department of Child and Adolescent Psychiatry, Pitié-Salpêtrière Hospital, Paris, France
[k]Psychiatry Department, EPS Barthélémy Durand, Etampes, France.
[l]Centre Hospitalier Universitaire Sainte-Justine, University of Montreal, Montreal, Quebec, Canada.
[m]Department of Child and Adolescent Psychiatry and Psychotherapy, University Medical Centre Göttingen, von-Siebold-Str. 5, 37075, Göttingen, Germany
[n]Technische Universität Dresden, Dresden, Germany
[o]Centre for Population Neuroscience and Stratified Medicine (PONS), Department of Psychiatry and Neuroscience, Charité Universitätsmedizin Berlin, Germany
[p]Department of Psychiatry and Psychotherapy CCM, Charité – Universitätsmedizin Berlin, Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany
[q]School of Psychology and Global Brain Health Institute, Trinity College Dublin, Ireland
[r]Institute for Science and Technology of Brain-inspired Intelligence (ISTBI), Fudan University, Shanghai, China
[s]Institute for Molecular Bioscience, the University of Queensland, Brisbane, Australia

ABSTRACT

Incomplete Hippocampal Inversion (IHI) is an atypical anatomical pattern of the hippocampus that has been associated with several brain disorders (epilepsy, schizophrenia). IHI can be visually detected on coronal T1 weighted MRI images. IHI can be absent, partial or complete (no IHI, partial IHI, IHI). However, visual evaluation can be long and tedious, justifying the need for an automatic method. In this paper, we propose, to the best of our knowledge, the first automatic IHI detection method from T1-weighted MRI. The originality of our approach is that, instead of directly detecting IHI, we propose to predict several anatomical criteria, which each characterize a particular anatomical feature of IHI, and that can ultimately be combined for IHI detection. Such individual criteria have the advantage of providing interpretable anatomical information regarding the morphological aspect of a given hippocampus. We relied on a large population of 2,008 participants from the IMAGEN study. The approach is general and can be used with different machine learning models. In this paper, we explored two different backbone models for the prediction: a linear method (ridge regression) and a deep convolutional neural network. We demonstrated that the interpretable, anatomical based prediction was at least as good as when predicting directly the presence of IHI, while providing interpretable information to the clinician or neuroscientist. This approach may be applied to other diagnostic tasks which can be characterized radiologically by several anatomical features.

Keywords: Deep Learning, Machine Learning, MRI, Incomplete Hippocampal Inversion

1. INTRODUCTION

Incomplete Hippocampal Inversions are found in around 20% of the general population and are more commonly observed in the left hemisphere (17.1% left hemisphere and 6.5% right hemisphere).[1] While their origin is not well understood, they are thought to occur during the pre-natal development of the temporal lobe. IHIs have been shown to have a higher prevalence in patients with epilepsy (30-50% of the population),[2–5] or schizophrenia.[6] This suggests that IHI might play a role in the development of several brain disorders, and more research is needed to investigate the association between IHI and other psychiatric or neurodevelopmental disorders.

However, IHIs need to be detected visually by a trained rater. This can be a long and tedious task and could greatly benefit from an automated method. We are not aware of any such method. IHIs can be absent, partial or total and they are thus associated with a global three-class score (no IHI, partial IHI, total IHI). However, this global score does not account for the different anatomical characteristics of IHI and is likely to lack reproducibility from one laboratory to another. Thus, different authors have proposed to rate IHI using a set of criteria.[1,3,4] Specifically, Cury et al.[1] proposed a rating scale composed of five criteria/dimensions assessing the different characteristics of the hippocampus: verticality/roundness, medial positioning and neighbouring sulci characteristics (sulcal depth). Individual criteria can be summed-up to form an IHI score representing the IHI level of a given hippocampus. The IHI criteria allow for a more specific assessment of individual characteristics and allow designing a more interpretable automatic rating method.

The aim of this work was to develop an automatic method to detect IHI from anatomical MRI. More specifically, we propose to predict individual anatomical criteria in order to obtain an intrinsically interpretable rating. The predicted criteria are subsequently combined to detect IHI. We then aimed to assess whether this prediction strategy leads to performances at least on par with those obtained when directly predicting the IHI status.

2. MATERIALS AND METHODS

2.1 Materials

We studied 2,008 participants from the IMAGEN study.[7] We included all participants with a T1-weighted anatomical MRI acquired at 3 Tesla. The average age at MRI was 14.5 years (range: 12.9 - 17.2). 51% participants were females, 49% males and sex information was missing for one. Both the global three-class criterion (denoted as C0) corresponding to IHI detection and the individual interpretable criteria (denoted from

Further author information: (Send correspondence to Lisa Hemforth)
Lisa Hemforth: E-mail: hemforthl@gmail.com

C1 to C5) were assessed on all MRI images by trained raters. The sum of individual criteria was called the IHI score denoted as SC.[1] Each of these criteria and scores have been evaluated separately on the left (_L) and on the right (_R) hemisphere. Local ethics committees approved the study. Participants as well as their parents gave informed written consent.

2.2 MRI pre-processing

We processed the MRI using the t1-volume pipeline implemented in Clinica.[8,9] This pipeline is a wrapper of the *Segmentation*, *Run Dartel* and *Normalise to MNI Space* routines implemented in SPM. First, the Unified Segmentation procedure[10] is used to simultaneously perform tissue segmentation, bias correction and spatial normalization of the input image. Next, a group template is created using DARTEL, an algorithm for diffeomorphic image registration,[11] from the participants' tissue probability maps on the native space, usually GM, WM and CSF tissues, obtained at the previous step. The DARTEL to MNI method[11] is then applied, providing a deformable registration of the native space images into the MNI space. We further cropped the gray-matter maps into a box of interest containing both hippocampi and the surrounding sulci.

2.3 Split between learning and testing set

We isolated 25% (502) of the participants to form a test set. We performed the split prior to running any analysis and only used the test set to evaluate results. This left a learning data-set of 1,506 participants (also used for model selection and hyperparameter optimization through cross-validation within these 1506 participants). We stratified the split based on all IHI criteria as well as age, weight, height, sex, handedness and imaging centre. In practice, we performed 200 random splits, and selected the one that minimised differences in distributions for all considered variables between the learning and test set (based on a Kolmogorov-Smirnoff test).

2.4 Proposed approach

The core of our approach is to predict the individual anatomical criteria C1 to C5, in place of the global criterion C0. The individual predicted criteria are then combined to detect IHI. The first criterion (C1) assesses the verticality and roundness of the hippocampal body. The second criterion (C2) evaluates the verticality and depth of the collateral sulcus. The third criterion (C3) quantifies the medial position of the hippocampus. The fourth criterion (C4) is a binary score defining if the subiculum is bulging upwards or not. However, we did not use this criterion because it leads to difficulties in human annotations and because it is normal in the overwhelming majority of participants ($> 97\%$[1]). The fifth criterion (C5) assesses whether any sulci of the fusiform gyrus exceed the level of the subiculum. Each of these criteria is rated on a 2 point scale with 0.5 steps for criteria 1 to 3 and 1 point steps for criterion 5. A schematic of these criteria extracted from the paper by Claire Cury et al. can be found in figure 1.

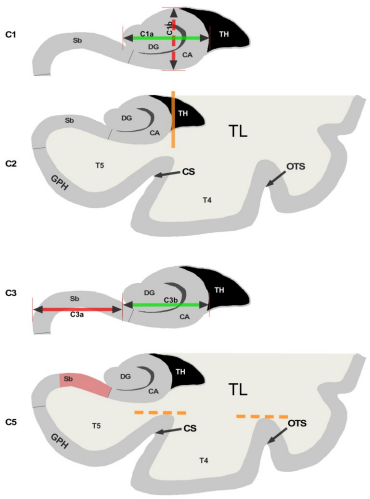

Figure 1. Schematic of criteria 1,2,3 and 5 reproduced from [1] (CC BY).

We trained machine learning models to predict each score (C1, C2, C3, C5) separately. Our approach can work with various machine learning models as a backbone. In this work, we compared two models. First, we considered a linear model (ridge regression) and used a cross-validation to estimate the best hyper-parameter. Next, we considered a convolutional neural network with five convolutional layers and three fully connected layers, denoted as Conv5-FC3 in the following, and implemented in ClinicaDL .[12,13] It was trained to perform regression using the mean squared error loss. The deep-learning model was trained over 50 epochs and we performed early stopping, i.e. the model with the lowest validation loss was used for further analysis. Note that none of these operations involved the test set in order to not to bias the results. Both the linear and deep learning models used as input the voxel-based gray-matter maps described in Section 2.2. Finally, we summed the predictions for the individual criteria and the result was denoted as SC_add_L (resp. SC_add_R) for the left (resp. right) hemisphere.

2.5 Comparison to the global criterion

For comparison, we trained the same models to predict the global criterion C0. We searched for the optimal threshold to discretize the continuous IHI score into three classes of C0. To that purpose, we iteratively searched for the threshold on SC that gave the most accurate classification for C0 (as measured by balanced accuracy). Absence of IHI (C0=0) corresponds to IHI scores below 2.25 , partial IHIs (C0=0.5) to scores between 2.25 and 4.25, total IHI (C0=1) to scores above 4.25. Note that we only used the learning set and the manually-obtained SC to compute these thresholds, in order not to bias the results. Finally, we also compared our approach to a direct prediction of SC.

2.6 Performance metrics and statistical analysis

To compare the prediction to the ground truth, we used 1) a quadratic weighted kappa score for the discretized SC, the global criterion C0 and the individual criteria C1, C2, C3 because these are discrete ordered variables; 2) a non-weighted kappa for C5 as it is a binary variable; 3) inter-class correlation coefficients (ICC) for the continuous SC. We performed a bootstrap on the isolated test set, using 1000 iterations of 502 samples, the same size as the isolated test set. From this, we deduced the mean kappa/weighted kappa/ICC score along with the standard error.

2.7 Visual analysis

For the ridge regression, we extracted a weight map, i.e. a 3D image showing the weight attributed to each voxel. We computed a saliency map[14] from the Conv5-FC3 models using the implementation provided in ClinicaDL.[13] To visualise which regions contribute most to the models' decisions, we only show the 100 voxels with the highest values over-layed on a T1 MRI image.

3. RESULTS

3.1 Evaluation of the proposed approach

Table 1 presents the performance obtained for the prediction of each individual criterion, when computing the sum of the predictions for C1, C2, C3 and C5 to obtain the IHI score, and when predicting the IHI score directly. The deep learning model systematically achieved higher performance compared to ridge regression. However, this difference ranged from small (about 0.05 points) to very high (about 0.25 points). Kappa and ICC scores were systematically lower in the right hemispheres, which we attributed to to the lower number of IHIs on this side.The fifth criterion was predicted with greater difficulty due to its unbalanced nature[1] (right side : 85%, 6%, 9%; left side : 59%, 20%, 20%). When summed, predictions obtained from individual criteria produced comparable results to predicting directly the IHI scores.

3.2 Comparison to the direct prediction of C0

Table 2 displays results for prediction of C0, either directly, or through thresholding the sum of the predictions of the individual criteria, or through thresholding the prediction of the IHI score (SC). Overall, the performance of the proposed approach (predicting interpretable individual scores) was comparable to the direct prediction of C0. For ridge regression, there were even some cases where the results were substantially better.

Table 1. Performances for the prediction of individual criteria C1, C2, C3, C5 and for the direct prediction of the IHI score (either using the sum of the predictions of individual criteria, denoted as "SC_{L,R}_add", or using direct prediction of SC, denoted as "SC_{L,R}"). The table shows the mean ± the standard error computed using bootstrap on the test set.

	C1_L	C1_R	C2_L	C2_R	C3_L	C3_R
	Weighted Kappa					
Ridge regression	0.558±0.027	0.256±0.037	0.675±0.024	0.599±0.019	0.717±0.019	0.627±0.031
Conv5-FC3	0.717±0.022	0.501±0.033	0.749±0.021	0.644±0.028	0.769±0.018	0.694±0.028

	C5_L	C5_R	SC_L_add	SC_R_add	SC_L	SC_R
	Kappa		ICC			
Ridge regression	0.268±0.031	0.187±0.034	0.684±0.023	0.576±0.031	0.708±0.021	0.633±0.024
Conv5-FC3	0.502±0.031	0.331±0.047	0.811± 0.015	0.678±0.032	0.788±0.015	0.703±0.027

Table 2. Results for prediction of C0, either directly, or through thresholding the sum of the predictions of the individual criteria ("SC_{L,R}_add"), or through thresholding the direct prediction of the IHI score ("SC_{L,R}"). Mean ± standard error (of each metric), computed using bootstrap on the test set.

	C0_L	C0_R	SC_L_add	SC_R_add	SC_L	SC_R
Ridge Regression	0.546±0.031	0.500±0.047	0.640±0.027	0.563±0.037	0.628±0.029	0.601±0.033
Conv5-FC3	0.763±0.026	0.668±0.045	0.739±0.023	0.647±0.037	0.740±0.023	0.610±0.043

3.3 Visual interpretation

Weight maps and saliency maps can be seen in Figure 2. We noticed that both models seem to rely on voxels located in the hippocampus and the surrounding gyri. The hippocampus seems to be outlined for C1 and C3 while the gyri are highlighted for C2 and C5. The saliency maps are slightly less clear and exhibit sparser results. However, a similar tendency may be observed.

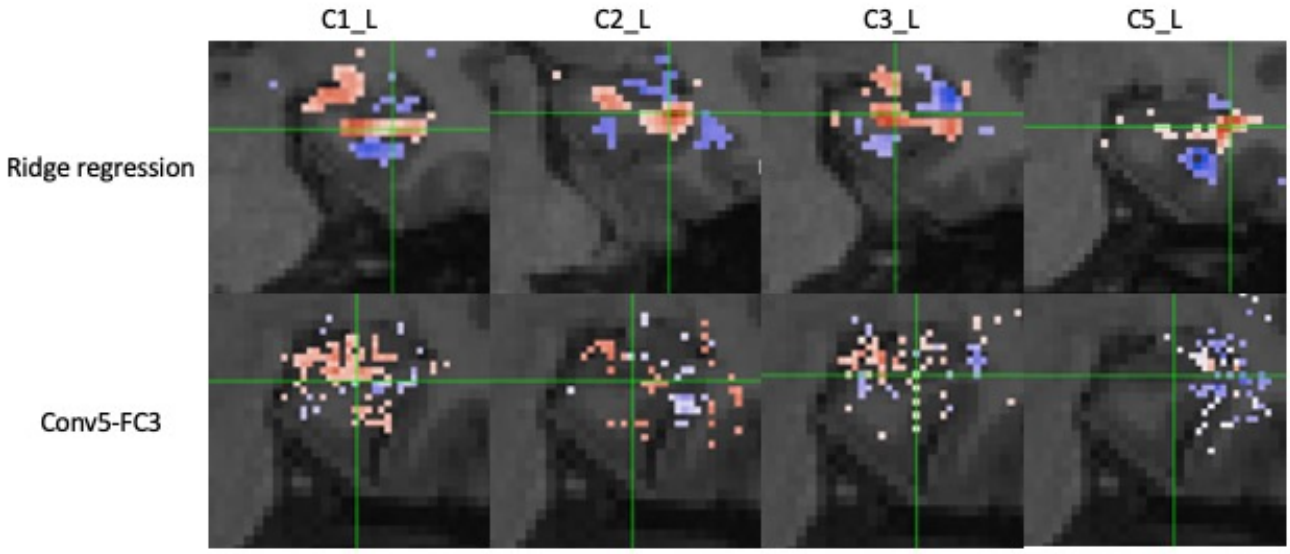

Figure 2. Weight maps extracted from ridge regression and saliency maps from the Conv5-FC3 for C1, C2, C3, and C5 in the left hemisphere on T1 weighted MRI image.

4. CONCLUSION

In this paper, we have proposed to automatically detect IHI by predicting anatomically interpretable individual criteria. We showed that this approach does not decrease predictive performance compared to directly predicting

the presence of IHI. Predicting individual criteria provides much more information about the specific anatomical characteristics underlying the IHI of a given participant, thereby providing more interpretable information to the clinician or neuroscientist. This training strategy has the potential to be applied to other diagnosis tasks which can be characterized by individual interpretable criteria.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the French government under the management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6). BCD is funded by Inria and CJ Martin fellowship funded by the N HMRC (app 1161356).

The Imagen study received support from the following sources: the European Union-funded FP6 Integrated Project IMAGEN (Reinforcement-related behaviour in normal brain function and psychopathology) (LSHM-CT-2007-037286), the Horizon 2020 funded ERC Advanced Grant 'STRATIFY' (Brain network based stratification of reinforcement-related disorders) (695313), Human Brain Project (HBP SGA 2, 785907, and HBP SGA 3, 945539), the Medical Research Council Grant 'c-VEDA' (Consortium on Vulnerability to Externalizing Disorders and Addictions) (MR/N000390/1), the National Institute of Health (NIH) (R01DA049238, A decentralized macro and micro gene-by-environment interaction analysis of substance use behavior and its brain biomarkers), the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, the Bundesministeriumfür Bildung und Forschung (BMBF grants 01GS08152; 01EV0711; Forschungsnetz AERIAL 01EE1406A, 01EE1406B; Forschungsnetz IMAC-Mind 01GL1745B), the Deutsche Forschungsgemeinschaft (DFG grants SM 80/7-2, SFB 940, TRR 265, NE 1383/14-1), the Medical Research Foundation and Medical Research Council (grants MR/R00465X/1 and MR/S020306/1), the National Institutes of Health (NIH) funded ENIGMA (grants 5U54EB020403-05 and 1R56AG058854-01), NSFC grant 82150710554 and European Union funded project 'environMENTAL', grant no: 101057429. Further support was provided by grants from: - the ANR (ANR-12-SAMA-0004, AAPG2019 - GeBra), the Eranet Neuron (AF12-NEUR0008-01 - WM2NA; and ANR-18-NEUR00002-01 - ADORe), the Fondation de France (00081242), the Fondation pour la Recherche Médicale (DPA20140629802), the Mission Interministérielle de Lutte-contre-les-Drogues-et-les-Conduites-Addictives (MILDECA), the Assistance-Publique-Hôpitaux-de-Paris and INSERM (interface grant), Paris Sud University IDEX 2012, the Fondation de l'Avenir (grant AP-RM-17-013), the Fédération pour la Recherche sur le Cerveau; the National Institutes of Health, Science Foundation Ireland (16/ERCD/3797), U.S.A. (Axon, Testosterone and Mental Health during Adolescence; RO1 MH085772-01A1) and by NIH Consortium grant U54 EB020403, supported by a cross-NIH alliance that funds Big Data to Knowledge Centres of Excellence.

REFERENCES

[1] Cury, C., Toro, R., Cohen, F., Fischer, C., Mhaya, A., Samper-González, J., Hasboun, D., Mangin, J.-F., Banaschewski, T., Bokde, A. L. W., Bromberg, U., Buechel, C., Cattrell, A., Conrod, P., Flor, H., Gallinat, J., Garavan, H., Gowland, P., Heinz, A., Ittermann, B., Lemaitre, H., Martinot, J.-L., Nees, F., Paillère Martinot, M.-L., Orfanos, D. P., Paus, T., Poustka, L., Smolka, M. N., Walter, H., Whelan, R., Frouin, V., Schumann, G., Glaunès, J. A., Colliot, O., and the Imagen Consortium, "Incomplete hippocampal inversion: A comprehensive MRI study of over 2000 subjects.," *Front Neuroanat.* **9**, 160 (2015).

[2] Lehéricy, S., Dormont, D., Sémah, F., Clémenceau, S., Granat, O., Marsault, C., and Baulac, M., "Developmental abnormalities of the medial temporal lobe in patients with temporal lobe epilepsy.," *AJNR Am J Neuroradiol.* **16**, 617–626 (1995).

[3] Baulac, M., De Grissac, N., Hasboun, D., Oppenheim, C., Adam, C., Arzimanoglou, A., Semah, F., Lehéricy, S., Clémenceau, S., and Berger, B., "Hippocampal developmental changes in patients with partial epilepsy: magnetic resonance imaging and clinical aspects.," *Ann Neurol.* **44**, 223–33 (1998).

[4] Bernasconi, N., Kinay, D., Andermann, F., Antel, S., and Bernasconi, A., "Analysis of shape and positioning of the hippocampal formation: an MRI study in patients with partial epilepsy and healthy controls.," *Brain.* **128**, 2442–2452 (2005).

[5] Bajic, D., Kumlien, E., Mattsson, P., Lundberg, S., Wang, C., and Raininko, R., "Incomplete hippocampal inversion - is there a relation to epilepsy?," *Eur Radiol.* **19**, 2544–2550 (2009).

[6] Roeske, M. J., McHugo, M., Vandekar, S., Blackford, J. U., Woodward, N. D., and Heckers, S., "Incomplete hippocampal inversion in schizophrenia: prevalence, severity, and impact on hippocampal structure.," *Mol Psychiatry.* **26**, 5407–5416 (2021).

[7] Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Büchel, C., Conrod, P. J., Dalley, J. W., Flor, H., Gallinat, J., Garavan, H., Heinz, A., Itterman, B., Lathrop, M., Mallik, C., Mann, K., Martinot, J.-L., Paus, T., Poline, J.-B., Robbins, T. W., Rietschel, M., Reed, L., Smolka, M., Spanagel, R., Speiser, C., Stephens, D. N., Ströhle, A., Struve, M., and the IMAGEN consortium, "The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology.," *Mol Psychiatry.* **15**, 1128–11390 (2010).

[8] Routier, A., Burgos, N., Díaz, M., Bacci, M., Bottani, S., El-Rifai, O., Fontanella, S., Gori, P., Guillon, J., Guyot, A., Hassanaly, R., Jacquemont, T., Lu, P., Marcoux, A., Moreau, T., Samper-Gonzalez, J., Teichmann, M., Thibeau-Sutre, E., Vaillant, G., and Colliot, O., "Clinica: An open-source software platform for reproducible clinical neuroscience studies.," *Frontiers in Neuroinformatics.* **15**, 1662–5196 (2021).

[9] Samper-González, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., Bertrand, A., Bertin, H., Habert, M.-O., Durrleman, S., Evgeniou, T., and Colliot, O., "Reproducible evaluation of classification methods in alzheimer's disease: Framework and application to MRI and PET data.," *NeuroImage* **183**, 504–521 (2018).

[10] John Ashburner, K. J. F., "Unified segmentation.," *NeuroImage* **26**, 839–851 (2005).

[11] Ashburner, J., "A fast diffeomorphic image registration algorithm.," *NeuroImage* **38**, 95–11 (2007).

[12] Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., and Colliot, O., "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," *Medical Image Analysis* **63**, 101694 (2020).

[13] Thibeau-Sutre, E., Diaz, M., Hassanaly, R., Routier, A., Dormont, D., Colliot, O., and Burgos, N., "ClinicaDL: an open-source deep learning software for reproducible neuroimaging processing," *Computer Methods and Programs in Biomedicine* **220**, 106818 (2022).

[14] Simonyan, K., Vedaldi, A., and Zisserman, A., "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint* **arXiv:1312.6034** (2013).