Simulation-based evaluation framework for deep learning unsupervised anomaly detection on brain FDG PET

Ravi Hassanaly, Simona Bottani, Benoît Sauty, Olivier Colliot, Ninon Burgos

Simulation-based evaluation framework for deep learning unsupervised anomaly detection on brain FDG PET

Ravi Hassanaly[a], Simona Bottani[a], Benoit Sauty[a], Olivier Colliot[a], and Ninon Burgos[a]

[a]Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France

ABSTRACT

Unsupervised anomaly detection using deep learning models is a popular computer-aided diagnosis approach because it does not need annotated data and is not restricted to the diagnosis of a disease seen during training. Such approach consists in first learning the distribution of anomaly free images. Images presenting anomalies are then detected as outliers of this distribution. These approaches have been widely applied in neuroimaging to detect sharp and localized anomalies such as tumors or white matter hyper-intensities from structural MRI. In this work, we aim to detect anomalies from FDG PET images of patients with Alzheimer's disease. In this context, the anomalies can be subtle and difficult to delineate, making the task more difficult and meaning that no ground truth exists to evaluate the approaches. We thus propose a framework to evaluate unsupervised anomaly detection approaches that consists in simulating realistic anomalies from images of healthy subjects. We demonstrate the use of this framework by evaluating an approach based on a 3D variational autoencoder.

Keywords: Unsupervised anomaly detection, Deep generative models, PET, Alzheimer's disease, Simulation framework

1. INTRODUCTION

As the global population gets older, we might face more cases of Alzheimer's disease (AD) and other types of dementia,[1] with an increase in social and financial costs.[2] Neurodegeneration caused by these diseases is visible on several imaging modalities,[3] including ^{18}F-fluorodeoxyglucose (FDG) PET in the form of hypo-metabolism.[4] By displaying subtle changes in brain metabolism, FDG PET can help the early diagnosis of dementia.

Recent breakthroughs in deep learning have offered many new possibilities in medical image analysis and algorithms are now able to accomplish complex tasks[5] such computer-aided diagnosis. A now widely used approach is unsupervised anomaly detection (UAD). The underlying idea is to learn the distribution of normal data and then detect potential out-of-distribution samples,[6] and thus identify abnormal cases. The first advantage of this method is that is does not require voxel-level annotation; another benefit is that the model should be able to detect any type of anomalies, without having seen them before.

One way of applying UAD to medical images is to train generative models such as variational autoencoders (VAE)[7] or generative adversarial networks (GAN)[8] to synthesize healthy looking images. Such model is only trained with images from subjects diagnosed as healthy. The assumption made is that if the model only learns to reconstruct healthy images, then the reconstruction of abnormal images will be inaccurate (and will ideally look healthy). The comparison of real and generated images should enable the detection and localization of pathological areas. We can distinguish two different objectives: the first one is to reconstruct an image that corresponds to the subject under investigation, and the second one is to generate images that look healthy.[9]

UAD has been widely applied to neuroimaging data to detect anomalies that are sharp and localized such as tumors or white matter hyper-intensities on structural MRI.[10] However, applying UAD to dementia is more challenging because the lesions, e.g., metabolic changes on FDG PET, are more diffuse and less intense. Moreover, there are no masks for anomalies to use as ground truth to evaluate UAD models. Therefore most of the studies

Further author information: (Send correspondence to Ravi Hassanaly)
Ravi Hassanaly: E-mail: ravi.hassanaly@icm-institute.org
Ninon Burgos: E-mail: ninon.burgos@cnrs.fr

rely on the residual error between the original image and its reconstruction,[11] which does not allow a precise evaluation of the model's capabilities.

We propose a new framework for the evaluation of UAD applied to dementia using synthetic data: we simulate anomalies on healthy images to generate a test set with pairs of diseased images and their healthy version. We can then reconstruct a pseudo-healthy version from the diseased image and compare it to the original healthy image that is the ground truth. We finally apply this framework to evaluate a new 3D VAE for UAD on FDG PET scans for dementia.

2. METHOD AND MATERIALS

2.1 Dataset

FDG PET scans used in this study were obtained from the ADNI database.[12,13] We selected images co-registered, averaged and uniformized to the same resolution to reduce the variability due to the use of different cameras. PET images were then processed using the `pet-linear` pipeline of the open-source Clinica[14] software: they were linearly registered to the standard MNI space, normalized in intensity using the average PET uptake in a region comprising cerebellum and pons, and cropped.

In ADNI, there is a total of 3511 FDG PET scans from 1600 participants. Since UAD models are trained only on healthy images, we selected the 301 cognitively normal (CN) subjects (733 images). We also selected the 311 baseline sessions of AD patients for testing purposes. All the other participants were discarded.

2.2 Simulation-based evaluation framework

The main challenges of pseudo-healthy reconstruction are to preserve the subject identity in the reconstructed image and to ensure that the model outputs are healthy brains.[9] To evaluate the quality of the reconstruction, four metrics are often used in the literature:[15] the mean absolute error (MAE), the mean squared error (MSE), the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM).[16] But if these metrics can be used to measure the conservation of the subject identity, they are not suited to evaluate healthiness of the synthesized images.

In our context, evaluating the healthiness of synthesized images would require a ground truth anomaly mask for each image or the intervention of a nuclear medicine physician who would qualitatively analyze the results of our model. However, we do not have any ground truth of the hypo-metabolism that we aim to detect and the expertise of a nuclear radiologist would be better exploited at a later stage of the evaluation process.

To tackle this issue, we generated a new test set by simulating hypo-metabolism on healthy images to have pairs of healthy (considered as ground truth) and diseased images. For this purpose, we designed a mask corresponding to regions associated with AD (parietal and temporal lobes)[17] that were extracted from the AAL3 atlas.[18] To obtain a realistic synthetic image, we smoothed the mask with a Gaussian convolution filter of $\sigma = 5$. We then reduced the intensity of the PET scan within the region defined by the mask by different factors to simulate various degrees of hypo-metabolism as we illustrate in Figure 1. Having such pairs of images allows comparing images reconstructed by the model from images presenting anomalies with their corresponding healthy images, hence better evaluating the model capacity to synthesize pseudo-healthy scans.

To ensure that the UAD model being evaluated can generalize to dementias other than AD, we generated masks corresponding to five other dementias: behavioral variant frontotemporal dementia (bvFTD), logopenic variant primary progressive aphasia (lvPPA), semantic variant PPA (svPPA), nonfluent variant PPA (nfvPPA) and posterior cortical atrophy (PCA) based on the regions defined by Burgos et al.[19] (Table 1).

2.3 3D variational autoencoder

To synthesize pseudo-healthy images, we proposed a 3D variational autoencoder that we trained on images from CN subjects to learn their distribution. A VAE is a deep probabilistic model[7] that aims to approximate the true distribution of the data with a simple parameterized distribution. To this end, all the samples $\mathbf{x}$ from the dataset will be projected through an encoder in a latent space of smaller dimension in which each sample $\mathbf{x}$ will be mapped to a Gaussian distribution $\mathcal{N}(\mu(\mathbf{x}), \sigma(\mathbf{x}))$. Then, the decoder will learn to reconstruct the input data

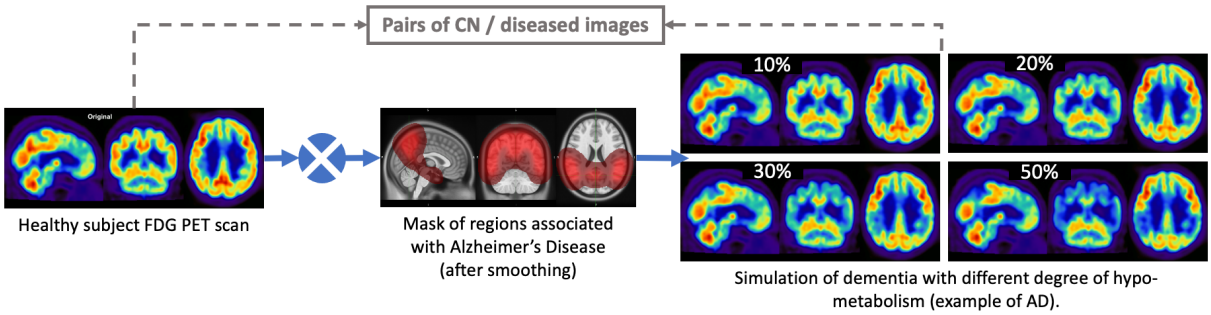


Figure 1: Simulation of images mimicking that of patients with dementia from healthy FDG PET scans. We use a mask of a region associated with dementia (here AD) and reduce the intensity of the voxels within this region after smoothing the mask. We can change the intensity of the hypo-metabolism to simulate dementia of various degrees.

from the latent representation z that will be sampled from this distribution. Let's note $\hat{\mathbf{x}}$ the reconstruction of $\mathbf{x}$, our loss function will be the sum of the L_2 loss and the Kullback-Leiber divergence between our distribution and the normal distribution $\mathcal{N}(0,1)$, which can be simplified in Equation 1 as

$$\mathcal{L}(\mathbf{x},\hat{\mathbf{x}}) = L_2(\mathbf{x},\hat{\mathbf{x}}) - \frac{1}{2}\left[\sigma(\mathbf{x})^2 + \mu(\mathbf{x})^2 - \log(\sigma(\mathbf{x})^2) - 1\right] \quad . \tag{1}$$

Even though some approaches using GANs have been developed in the literature,[9,20,21] we choose to use a variational autoencoder because it is simpler to train and does not rely only on residual error, but is also a likelihood-based model.[22] Moreover, variational autoencoders have the advantage of having a consistent latent space that can be used for further experimentation.

2.4 Experimental settings

We split our dataset of 301 CN subjects into training, validation and test sets at the subject's level to avoid any form of data leakage.[23] The split is stratified by sex and age to reduce biases (Table 2). 30 CN subjects compose the test set that is used to assess whether the healthy images are reconstructed as healthy. Then within the remaining CN subjects, 34 subjects belong to the validation set to monitor the training and 237 subjects are used to train our models. This represents 563 images for the training phase. In addition to CN subjects, we use the images of 311 AD patients acquired at baseline to create a second test set.

The encoder of the 3D VAE we implemented is composed of three convolutional layers and one dense layer, and the decoder is symmetrical. We used batch normalization after each convolutional layer and a leaky-ReLU activation function after each normalization. We empirically chose a latent space of size 128. The model was trained on 300 epochs, with a learning rate of 10^{-5} using the open-source ClinicaDL[24] software. Note that the images were down-sampled to a $80 \times 96 \times 80$ voxel size in order to reduce their dimension and the memory needed, which allowed increasing the batch size to 24.

3. RESULTS

3.1 Results on simulated AD-like FDG PET images

To evaluate the impact of the anomaly severity, we simulated different degrees of hypo-metabolism from 5% to 70% (Figure 2). We see that the MSE between the input and the output images is higher for more severe anomalies. This is expected as it means that the model cannot reconstruct well highly abnormal areas, and so should be able to detect them. We compared the MSE obtained on simulated data to that obtained when feeding images from the CN test set as inputs, which contain no anomaly. We observe that for low degree hypo-metabolism (<20%), the MSE are similar to that obtained for the CN test set. This means that the residual error due to the model imperfect reconstruction dissimulates the reconstruction error due to low degree anomalies.[22]

To confirm our observations, we computed a t-test assessing whether there was a significant difference in MSE between using healthy images as inputs and using images with various degrees of anomalies. The p-values

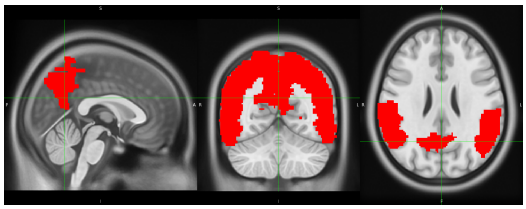
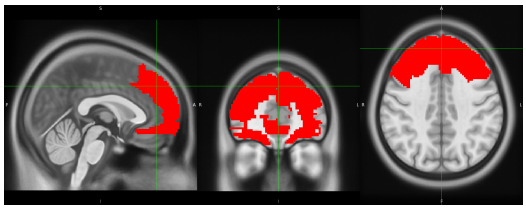
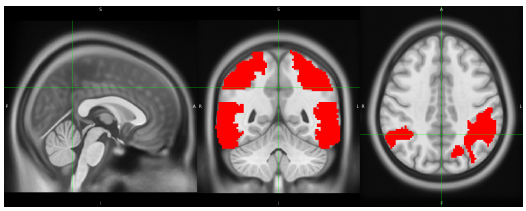
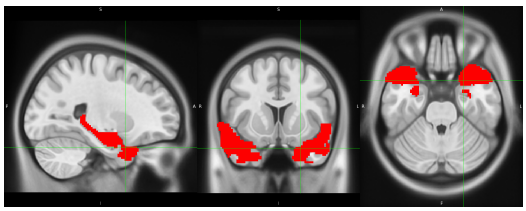
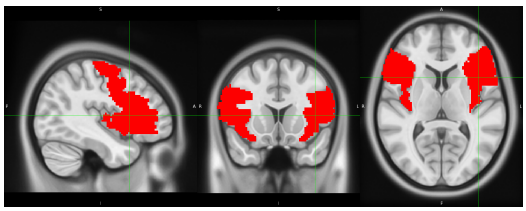
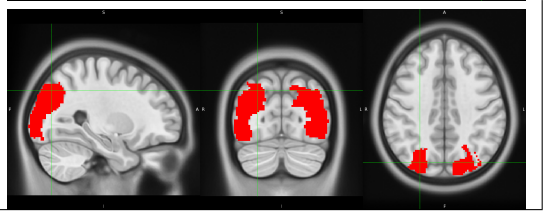
Dementias	Regions Associated	Masks
Alzheimer's disease (AD)	• **temporal lobe**, including the lateral and medial regions and temporal pole • **parietal lobe**, including the superior and inferior regions	
Behavioural variant frontotemporal dementia (bvFTD)	• **orbitofrontal region**, comprising the anterior, posterior, medial, and lateral orbital gyri, • **dorsolateral prefrontal region**, comprising the inferior, middle, and superior frontal gyri • **ventromedial prefrontal region**, comprising the gyrus rectus, medial frontal cortex, subcallosal area, and superior frontal gyrus medial segment.	
Logopenic variant primary progressive aphasia (lvPPA)	• **tempoparietal region**, comprising the inferior parietal lobule, posterior middle and superior temporal gyri.	
Semantic variant primary progressive aphasia (svPPA)	• **anterior temporal region**, comprising the hippocampus, amygdala and temporal pole.	
Nonfluent variant primary progressive aphasia (nfvPPA)	• **frontal region**, comprising the inferior frontal gyrus, precentral gyrus and anterior insula.	
Posterior cortical atrophy (PCA)	• **occipital region**, comprising the inferior, middle and superior occipital gyri.	

Table 1: Regions associated with different dementia as defined in Burgos et al[19] and the masks used for hypo-metabolism simulation.

were corrected for multiple comparisons using the Bonferroni method (the difference is statistically significant when p-value $< 0.05/9 = 0.0056$). The difference in MSE becomes significant for anomalies of degree 25 % and

Subset	Mean age ± std.	F %	M %
AD test	74.9 ± 7.7	41.5%	58.5%
CN test	74.4 ± 5.9	46.6%	53.4%
Train	74.5 ± 6.4	51.2%	48.8%

Table 2: Stratified splits between test and training sets: we try to keep the same age and sex distribution over different subsets to reduce bias.

above. This corroborates the results of Landau et al.[17] who defined that, on average in the ADNI dataset, the difference in metabolism between CN subjects and AD patients is ∼25% in a region of interest relevant to AD. Even though this might not readily translate to clinical application, it is a promising result as patients in the ADNI database are often at an early stage of the disease.

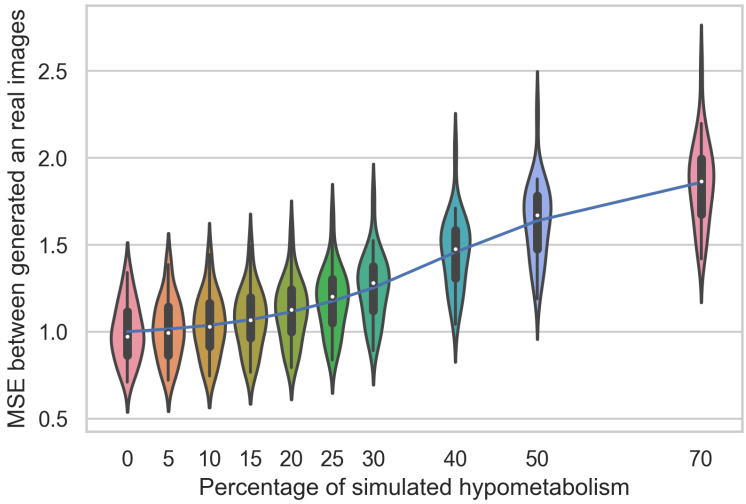


Figure 2: Evolution of the MSE with increasing degrees of hypo-metabolism simulating AD-like anomalies. Each MSE is normalized by the average MSE obtained when reconstructing from the original healthy images.

In Figure 3, we plotted the real image of a CN subject and its pseudo-healthy reconstruction, and the simulated AD version (with a hypo-metabolism degree of 30%) and its reconstruction for the same CN subject, together with the residual images. We observe that the input and output images of the CN subject are quite similar. The differences are due to the model imperfect reconstruction and correspond to the minimal error that it can achieve. When feeding the simulated hypo-metabolic image to the model, we observe that the reconstructed image looks healthier than the input image. The areas highlighted in the residual map correspond to the regions where hypo-metabolism was simulated. An interesting point also is that both images reconstructed from the same CN subject (either from the original image or the simulated hypo-metabolic one) are almost identical with a SSIM of 0.993. This shows that the model reconstructs the same image for the same subject whether the input image is healthy or presents anomalies.

3.2 Results when simulating various types of dementia

In this section, the degree of hypo-metabolism is set to 30% but the brain region where it is simulated changes to reflect various types of dementia. We report in Table 3 the different reconstruction metrics computed between the original PET scans from CN subjects in the test set and the images reconstructed from the hypo-metabolic scans simulating the different types of dementia. We observe that the metrics are similar for all the simulated dementias.

We also computed the metrics between the images reconstructed from the original healthy scans and the images reconstructed from the simulated hypo-metabolic scans. Both reconstructions are almost identical with

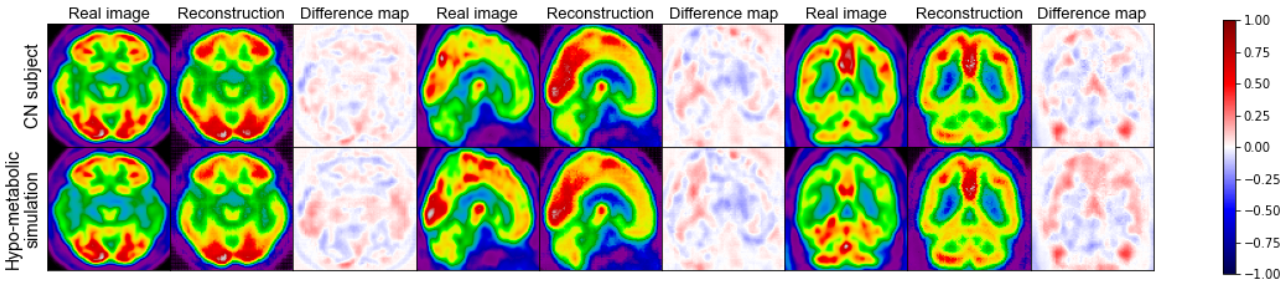


Figure 3: Example of results obtained from a real image of a CN subject (top row) and an image simulating AD hypo-metabolism based on the same CN subject (bottom row). For each plane, the first image is the input, the second one the model's reconstruction and the third one the difference (reconstruction - input).

	MSE	MAE	PSNR	SSIM
AD	0.00356 ± 0.00112	0.0414 ± 0.0059	24.6 ± 1.1	0.729 ± 0.056
bvFTD	0.00313 ± 0.00113	0.0388 ± 0.0062	25.2 ± 1.2	0.741 ± 0.056
lvPPA	0.00320 ± 0.00111	0.0392 ± 0.0061	25.1 ± 1.1	0.739 ± 0.055
svPPA	0.00293 ± 0.00118	0.0375 ± 0.0064	25.5 ± 1.3	0.749 ± 0.057
nfvPPA	0.00319 ± 0.00125	0.0388 ± 0.0066	25.2 ± 1.2	0.744 ± 0.057
PCA	0.00306 ± 0.00115	0.0381 ± 0.0063	25.4 ± 1.2	0.749 ± 0.056

Table 3: Reconstruction metrics computed between the original healthy PET scans from CN subjects in the test set and the images reconstructed with the 3D VAE from the hypo-metabolic scans simulating different types of dementia (average ± std.).

an SSIM on average superior to 0.99. We can conclude from this experiment that the model is able to reconstruct the healthy version of an image independently of the nature of the dementia that causes the anomaly.

3.3 Results on real images from the ADNI dataset

As final experiment, we compared the reconstructions obtained from the CN subjects from the test set and the AD patients. We can see in Table 4 that the reconstruction is on average better for CN subjects than AD patients, which is what we expect. In Figure 4, we also plotted the real image and the pseudo-healthy reconstruction for an AD patient and a CN subject from ADNI. We can observe that reconstruction quality is satisfactory for a very simple 3D model as we can recognize the subject in the output image. If we look at the AD patient, it seems like the model corrects the hypo-metabolism that we can observe in the PET scan, which is particularly visible on the bottom left corner of the axial slice and the left part of the coronal slice.

	MSE	MAE	PSNR	SSIM
AD test	0.00408 ± 0.00356	0.0433 ± 0.0116	24.5 ± 1.8	0.709 ± 0.101
CN test	0.00292 ± 0.00119	0.0373 ± 0.0065	25.6 ± 1.3	0.749 ± 0.057

Table 4: Reconstruction metrics obtained for real images of CN subjects and AD patients from ADNI on the test sets only (average ± std.).

4. DISCUSSION & CONCLUSION

The framework we proposed for UAD evaluation consists in simulating different types of dementia from a healthy PET image by reducing the intensity within a mask corresponding to regions known to be affected by the disease. This allows us to obtain pairs of healthy ground truth and abnormal images to evaluate UAD models. This evaluation framework can be used in future work to make an evaluation of the different generative models that are developed for dementia UAD on brain FDG PET.

We simulated with this method different degrees of anomalies and different types of dementia to evaluate an UAD approach in different conditions. We could first show that the model tested can detect anomalies of

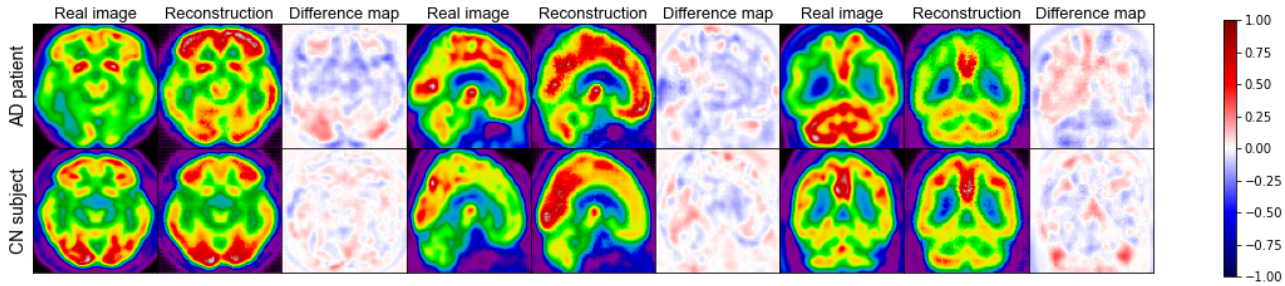


Figure 4: Example of results obtained from a real image of an AD patient (top row), a real image of a CN subject (bottom row). For each plane, the first image is the input, the second one the model's reconstruction and the third one the difference (reconstruction - input).

similar severity as found in a real AD dataset. We also showed that the model tested has similar results on other simulated dementias than AD, which shows the ability of the model to generalize to new diseases. We could also see that the images reconstructed from a healthy image or its simulated hypo-metabolic version are almost identical, which means that the model reconstructs a healthy version of the diseased image. We could improve this approach by adding more variability in the simulated diseased images by choosing sub-regions or randomly sampling the severity of the anomaly.

Finally, we showed that the simple 3D VAE we proposed leads to acceptable results in terms of reconstruction even though the results could be improved. We could enhance the VAE reconstruction capability by choosing a better posterior distribution approximation, adding a discriminator to the VAE to sharpen its output, or trying other autoencoder based models such as the adversarial autoencoder.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the French government under 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).

The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early AD.

REFERENCES

[1] Hebert, L. E., Weuve, J., Scherr, P. A., and Evans, D. A., "Alzheimer disease in the united states (2010–2050) estimated using the 2010 census," *Neurology* **80**(19), 1778–1783 (2013).

[2] Brookmeyer, R., Johnson, E., Ziegler-Graham, K., and Arrighi, H. M., "Forecasting the global burden of alzheimer's disease," *Alzheimer's & Dementia* **3**(3), 186–191 (2007).

[3] Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Feldman, H. H., Frisoni, G. B., Hampel, H., Jagust, W. J., Johnson, K. A., Knopman, D. S., Petersen, R. C., Scheltens, P., Sperling, R. A., and Dubois, B., "A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers," *Neurology* **87**(5), 539–547 (2016).

[4] Herholz, K., "FDG PET and differential diagnosis of dementia," *Alzheimer Disease and Associated Disorders* **9**(1), 6–16 (1995).

[5] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I., "A survey on deep learning in medical image analysis," *Medical image analysis* **42**, 60–88 (2017).

[6] Bulusu, S., Kailkhura, B., Li, B., Varshney, P. K., and Song, D., "Anomalous example detection in deep learning: A survey," *IEEE Access* **8**, 132330–132347 (2020).

[7] Kingma, D. P. and Welling, M., "Auto-encoding variational bayes," (2014).

[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., "Generative adversarial nets," in [*Advances in Neural Information Processing Systems*], **27** (2014).

[9] Xia, T., Chartsias, A., and Tsaftaris, S. A., "Adversarial pseudo healthy synthesis needs pathology factorization," in [*International Conference on Medical Imaging with Deep Learning*], 512–526, PMLR (2019).

[10] Baur, C., Denner, S., Wiestler, B., Navab, N., and Albarqouni, S., "Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study," *Medical Image Analysis* **69**, 101952 (2021).

[11] Choi, H., Ha, S., Kang, H., Lee, H., Lee, D. S., and Alzheimer's Disease Neuroimaging Initiative, "Deep learning only by normal brain PET identify unheralded brain anomalies," *EBioMedicine* **43**, 447–453 (2019).

[12] Jagust, W. J., Bandy, D., Chen, K., Foster, N. L., Landau, S. M., Mathis, C. A., Price, J. C., Reiman, E. M., Skovronsky, D., and Koeppe, R. A., "The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core," *Alzheimer's & Dementia* **6**(3), 221–229 (2010).

[13] Jagust, W. J., Landau, S. M., Koeppe, R. A., Reiman, E. M., Chen, K., Mathis, C. A., Price, J. C., Foster, N. L., and Wang, A. Y., "The Alzheimer's Disease Neuroimaging Initiative 2 PET Core: 2015," *Alzheimer's & Dementia* **11**(7), 757–771 (2015).

[14] 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-González, J., Teichmann, M., Thibeau-Sutre, E., Vaillant, G., Wen, J., Wild, A., Habert, M.-O., Durrleman, S., and Colliot, O., "Clinica: An open-source software platform for reproducible clinical neuroscience studies," *Frontiers in Neuroinformatics* **15** (2021).

[15] Nečasová, T., Burgos, N., and Svoboda, D., "Validation and evaluation metrics for medical and biomedical image synthesis," in [*Biomedical Image Synthesis and Simulation*], Burgos, N. and Svoboda, D., eds., 573–600, Elsevier (2022).

[16] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing* **13**(4), 600–612 (2004).

[17] Landau, S. M., Mintun, M. A., Joshi, A. D., Koeppe, R. A., Petersen, R. C., Aisen, P. S., Weiner, M. W., and Jagust, W. J., "Amyloid deposition, hypometabolism, and longitudinal cognitive decline," *Annals of Neurology* **72**(4), 578–586 (2012).

[18] Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J., and Joliot, M., "Automated anatomical labelling atlas 3," *Neuroimage* **206**, 116189 (2020).

[19] Burgos, N., Cardoso, M. J., Samper-González, J., Habert, M.-O., Durrleman, S., Ourselin, S., and Colliot, O., "Anomaly detection for the individual analysis of brain PET images," *Journal of Medical Imaging* **8**(2), 024003 (2021).

[20] Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G., "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in [*Information Processing in Medical Imaging*], *LNCS* (2017).

[21] Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., and Schmidt-Erfurth, U., "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Medical Image Analysis* **54** (2019).

[22] Meissen, F., Wiestler, B., Kaissis, G., and Rueckert, D., "On the pitfalls of using the residual as anomaly score," in [*Medical Imaging with Deep Learning*], (2021).

[23] 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).

[24] Thibeau-Sutre, E., Díaz, 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** (2022).