Unsupervised anomaly detection in 3D brain FDG PET: A benchmark of 17 VAE-based approaches

Ravi Hassanaly, Camille Brianceau, Olivier Colliot, Ninon Burgos

Unsupervised anomaly detection in 3D brain FDG PET: A benchmark of 17 VAE-based approaches

Ravi Hassanaly[1], Camille Brianceau[1], Olivier Colliot[1], and Ninon Burgos[1]

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

Abstract. The use of deep generative models for unsupervised anomaly detection is an area of research that has gained interest in recent years in the field of medical imaging. Among all the existing models, the variational autoencoder (VAE) has proven to be efficient while remaining simple to use. Much research to improve the original method has been achieved in the computer vision literature, but rarely translated to medical imaging applications. To fill this gap, we propose a benchmark of fifteen variants of VAE that we compare with a vanilla autoencoder and VAE for a neuroimaging use case relying on a simulation-based evaluation framework. The use case is the detection of anomalies related to Alzheimer's disease and other dementias in 3D FDG PET.
We show that among the fifteen VAE variants tested, nine lead to a good reconstruction accuracy and are able to generate healthy-looking images. This indicates that many approaches developed for computer vision applications can generalize to the unsupervised detection of anomalies of various shapes, intensities and locations in 3D FDG PET. However, these models do not outperform the vanilla autoencoder and VAE.

Keywords: Variational autoencoder · Deep generative models · Unsupervised anomaly detection · PET · Alzheimer's disease

1 Introduction

Recent advances in medical image analysis have allowed the emergence of algorithms that can perform complex tasks such as computer-aided diagnosis [7,10] with pseudo-healthy reconstruction for unsupervised anomaly detection (UAD). Contrary to supervised approaches, UAD does not require human annotations that are costly and time consuming, and enables the detection of any type of anomalies, without having seen them before. Most approaches rely on generative models to reconstruct healthy looking images, also called pseudo-healthy images [1,7,10]. The assumption is that if a model is trained with images from subjects diagnosed as healthy, the reconstruction of images with a pathology should not contain pathology-specific features and look like a healthy image. Comparing the pseudo-healthy reconstruction with the real image then allows the detection of anomalies.

The application context of our work is the detection of metabolic changes visible in brain ^{18}F-fluorodeoxyglucose (FDG) positron emission tomography

(PET) caused by Alzheimer's disease and other dementias [8]. These subtle changes appear several years before the first symptoms and can be used for early diagnosis [16]. In neuroimaging, deep learning methods for UAD have not been much applied for the diagnosis of dementia [9]. It is a challenging task because the metabolic abnormalities are diffuse and little intense, which makes them difficult to detect [3].

The different pseudo-healthy reconstruction approaches that have been developed for medical imaging rely on variational autoencoders (VAEs) [19], generative adversarial networks (GANs) [12] and more recently diffusion models [15]. We aim to compare VAE-based models as they have shown their efficacy for UAD in medical imaging [1,7], are easy to train, easily scalable, with good interpretation capacity thanks to their regularized latent space, and are able to handle small datasets. Much research to improve the original VAE has been achieved in the computer vision literature [2,5,11,14,18,21,22,23,25,27,29,30,32,36], but only a few have been translated to medical imaging applications [1,6,9,24,31].

We propose a benchmark of seventeen VAE-based models and show results in the context of pseudo-healthy reconstruction for dementia from 3D FDG PET. As far as we know, the only study that has compared VAEs for neuroimaging data is that of Baur et al. [1]. However, it was restricted to models that had already been used for medical imaging applications. Many other VAE extensions have thus not been assessed. Also, it was dedicated to the detection of very sharp and intense anomalies, such as brain tumors or multiple sclerosis lesions, which is very different from the identification of subtle anomalies found in PET images of patients with cognitive disorders. Finally, it was performed in 2D. Our work aims to contribute to this effort by evaluating a much wider set of approaches, including many that were never used in medical imaging, relying on the work of Chadebec et al. [4]. This will provide an insight into the performance that such models can achieve in detecting anomalies in 3D data when trained with a relatively small dataset (few hundreds of images) compared to most datasets used in the computer vision literature (several tens of thousands images). The models will be evaluated and compared based on reconstruction quality and on their ability to generate healthy looking images using a previously proposed simulation framework [13].

2 Methods

2.1 Variational autoencoder framework for pseudo-healthy image reconstruction

Let D be a set of medical images of the same modality acquired following a similar protocol. D can contain healthy and pathological images and can be divided in respectively two complementary subsets D_h and D_p. Let's take as an example a set of FDG PET images $\mathbf{x} \in D_h$ whose distribution is $p(\mathbf{x})$. The goal of pseudo-healthy image reconstruction is to generate an FDG PET image of healthy appearance. The idea is to approximate the healthy image true distribution $p(\mathbf{x})$ with a chosen model $p_\theta(\mathbf{x})$ such that $p_\theta(\mathbf{x}) \approx p(\mathbf{x})$. Then, during reconstruction,

the images (of healthy subjects or patients) are projected into that "healthy images" learned subspace by the generative model.

This can be modeled using the VAE framework [19] by assuming that a latent variable $\mathbf{z}$ is involved in the generation process of $\mathbf{x}$: $p_\theta(\mathbf{x}) = \int_z p(\mathbf{z}) p_\theta(\mathbf{x} \mid \mathbf{z}) \mathrm{d}z$ where $\mathbf{z} \sim p_\theta(\mathbf{z})$ is the prior distribution on the latent space and $p_\theta(\mathbf{x} \mid \mathbf{z})$ is the generative model (or the decoder) that learns to generate healthy images from $\mathbf{z}$. To compute the appropriate $\mathbf{z}$ for each data input $\mathbf{x}$ of our dataset, we need the posterior distribution $p_\theta(\mathbf{z} \mid \mathbf{x})$. Since it is untractable, we approximate it using variational inference by introducing another model $q_\phi(\mathbf{z} \mid \mathbf{x})$ such that $q_\phi(\mathbf{z} \mid \mathbf{x}) \approx p_\theta(\mathbf{z} \mid \mathbf{x})$. $q_\phi(\mathbf{z} \mid \mathbf{x})$ is the inference model (or encoder). Both the decoder and encoder are parametric models whose parameters are given by a neural network.

The objective is to maximize the likelihood of $p_\theta(\mathbf{x})$, which is equivalent to maximizing the evidence lower bound, which defines our loss function $\mathcal{L}_{\theta,\Phi}$ [20]

$$\log\left(p_\theta(\mathbf{x})\right) \geq \mathcal{L}_{\theta,\Phi}(x) = \mathbb{E}_{q_\Phi(\mathbf{z}|\mathbf{x})}\Big[\log\big(p_\theta(\mathbf{x} \mid \mathbf{z})\big)\Big] - D_{\mathrm{KL}}\Big(q_\Phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\Big) \quad (1)$$

with D_{KL} the Kullback-Leibler divergence.

During the training process, we learn an approximation of the posterior distribution $q_\phi(\mathbf{z} \mid \mathbf{x})$ for $x \in D_h$ as we train our model using only healthy subjects. When using the model for inference, we use this approximate posterior to estimate the latent variable $\mathbf{z}$ for $\mathbf{x} \in D$ (it can be from D_h or D_p).

2.2 Extensions to the variational autoencoder framework

As explained in detail in [4], several contributions have been proposed to improve the VAE framework. They can be divided into four categories that correspond to different objectives:

- improve the prior distribution $p(\mathbf{z})$ by using a variational mixture of posteriors as prior (VAMP) [30], by learning the prior on a discrete latent space with vector quantized-VAE (VQVAE) [32], or by substituting the prior with a density estimation method using regularization with a gradient penalty (RAE-GP), or an ℓ^2 penalty on the decoder (RAE-ℓ^2) [11];
- better estimate the lower bound by using importance weighting (IWAE) [2], and using a linear normalizing flow (VAE-lin-NF) [25] or an inverse autoregressive flow (VAE-IAF) [21] to better estimate the posterior;
- encourage disentanglement of the features in the latent space by adding a weight to balance the terms of the loss in Eq. 1 (β-VAE) [14], decomposing the loss to show a total correlation term (β-TC VAE) [5], or by encouraging the distribution of the latent variable $q(\mathbf{z})$ to be factorial (FactorVAE) [18];
- and change the distance computed between the distributions by adding the mutual information between $\mathbf{x}$ and $\mathbf{z}$ as regularization (InfoVAE) [36], using another divergence term in the loss such as the maximum mean discrepancy in the Wasserstein autoencoder (WAE) [29] or a discriminator to differentiate a prior's sample from a posterior's sample in the adversarial autoencoder

(AAE) [23], or by changing the reconstruction metric for another similarity metric such as the multi-scale structural similarity (MSSSIM-VAE) [27], or for the prediction of a discriminator on the output of the VAE (VAEGAN) [22]. In our benchmark, these models will be compared to the autoencoder (AE) and VAE [19], which makes a total of seventeen models. All of these methods have shown great results in other fields of computer vision, and, since VAE-based models can learn the data distribution on a small dataset, we keep the focus on them and aim to assess their performance in the context of medical imaging.

2.3 Evaluation of the models

We can distinguish two main objectives when generating pseudo-healthy images: preserving the subject's identity in the reconstructed image and ensuring that the reconstruction appears healthy [35].

For the subject identity preservation, we evaluate the models on real images from healthy subjects only: the pseudo-healthy reconstruction of an image of a healthy subject should be identical to the input. This is assessed using three commonly used paired reconstruction metrics: the mean-squared error (MSE), the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) [33].

To evaluate the capability of each model to reconstruct healthy looking images, since we do not have access to ground-truth lesions masks, we use the evaluation framework that has been introduced in [13]. It consist in simulating the effect of the disease by reducing the intensity of the PET uptake within regions associated with different dementias, thus mimicking regional hypometabolism [3]. After locally reducing the intensity of the image by a certain percentage, a Gaussian smoothing is applied to have a realistic result and diffuse anomalies. That way we can have pairs of diseased images with the original healthy scan that is used as ground-truth for the pseudo-healthy reconstruction as we do not have ground truths for images from real patients in our dataset. We simulate five different dementias on images of healthy subjects: Alzheimer's disease (AD), behavioral variant frontotemporal dementia (bvFTD), logopenic variant primary progressive aphasia (lvPPA), semantic variant PPA (svPPA) and posterior cortical atrophy (PCA). This allows us to evaluate the capability of the model to generalize to anomalies caused by different dementia subtypes. In addition, we simulate different degrees of AD severity by varying the reduction in intensity from five to seventy percents to study the sensitivity of the UAD approaches on subtle and severe anomalies. We compute the reconstruction error in the whole image, in the region associated with the simulated dementia and in the complementary of this region in the brain.

2.4 Materials

FDG PET scans used in this study were obtained from the publicly available ADNI database [17] (`https://adni.loni.usc.edu`). We selected FDG PET images co-registered, averaged and uniformized to a resolution of 8 mm FWHM to reduce the variability due to the use of different scanners. The images were

then linearly registered to the standard MNI space, normalized in intensity using the average PET uptake in a region comprising cerebellum and pons, and cropped using the Clinica [26] `pet-linear` pipeline. We finally down-sampled the images to a voxel size of $80 \times 96 \times 80$ to reduce their dimension and the memory usage.

ADNI includes a total of 733 FDG PET scans of cognitively normal (CN) participants with a stable diagnosis over a three-year window (corresponding to 301 subjects). We discarded 144 images that were not correctly registered according to the quality check algorithms implemented in ClinicaDL [28].

2.5 Experimental setting

We split our dataset of 247 remaining CN subjects at the subject's level to avoid data leakage [34]: 50 CN subjects (50 images) compose the test set, 19 subjects (19 images) belong to the validation set and 178 subjects (452 images) are used to train our models. The split is stratified by sex and age to reduce biases. The 50 images of the CN subjects from the test set are also used to simulate the hypometabolic images mimicking various dementias and AD severity degrees.

For the comparison to be as fair as possible, all the models share the same encoder and decoder architecture. The encoder is composed of three blocks that are the succession of a 3D convolutional layer and a batch normalization with a ReLU activation. Then the tensor is flatten and passes through a dense layer to output a one dimensional latent space. The decoder is almost symmetrical: it is composed of a dense layer followed by three blocks that are composed of a 3D deconvolutional layer and a batch normalization with a leaky ReLU activation. We tested several sizes of latent space (16, 64, 128 and 256), but as we observed similar performance, we report the results for a size of 128, consistent with the choice made in [1].

We also use the same training parameters and environment to train all the models. We trained each model on 300 epochs with a learning rate of 10^{-5} and a batch size of 24 on a HPC with Nvidia Tesla V100 GPUs that have 32GB of memory. We are aware that model performance can greatly vary depending on these parameters, but for fair comparison we decided to choose the best parameters on the VAE and use the same for all models. It takes on average between 1' and 1'30" to train one epoch with comparable performance for each model on our computer cluster, meaning around 7 h per model for 300 epochs.

VAE-based model implementation relies on Pythae [4] and neuroimage processing on ClinicaDL [28], two open source software tools. The code used for this study is available on GitHub and can be used to reproduce the experiments: `https://github.com/ravih18/VAE-models-for-UAD`.

3 Results

3.1 Pseudo-healthy reconstruction from images of control subjects

We first assessed whether the different models could preserve the subject's identity by computing the MSE, PSNR and SSIM between the input and reconstructed

Table 1. Reconstruction metrics computed between the pseudo-healthy reconstructions obtained with the various models evaluated and the original healthy PET image of CN subjects from the test set. Light gray highlights the worst performing models.

Model	MSE ↓	PSNR (dB) ↑	SSIM ↑
AE	0.02694 ± 0.00603	25.78 ± 0.84	0.725 ± 0.033
VAE [19]	0.02471 ± 0.00517	26.15 ± 0.79	0.771 ± 0.027
VAMP [30]	1.09029 ± 0.10416	9.64 ± 0.41	0.057 ± 0.015
RAE-GP [11]	0.02363 ± 0.00480	26.34 ± 0.79	0.750 ± 0.030
RAE-ℓ^2 [11]	0.02385 ± 0.00532	26.31 ± 0.83	0.761 ± 0.029
VQVAE [32]	0.02645 ± 0.00608	25.87 ± 0.85	0.731 ± 0.032
IWAE [2]	0.03531 ± 0.00711	24.60 ± 0.80	0.692 ± 0.030
VAE-lin-NF [25]	0.12887 ± 0.02875	18.99 ± 0.89	0.483 ± 0.036
VAE-IAF [21]	0.02900 ± 0.00560	25.45 ± 0.77	0.706 ± 0.032
β-VAE [14]	0.03927 ± 0.00654	24.12 ± 0.71	0.708 ± 0.028
β-TC VAE [5]	0.02819 ± 0.00499	25.55 ± 0.67	0.729 ± 0.031
FactorVAE [18]	0.02869 ± 0.00550	25.49 ± 0.74	0.704 ± 0.032
InfoVAE [36]	0.03223 ± 0.00566	24.97 ± 0.69	0.706 ± 0.030
WAE [29]	0.02920 ± 0.00509	25.40 ± 0.66	0.690 ± 0.032
AAE [23]	0.02919 ± 0.00597	25.43 ± 0.81	0.709 ± 0.032
MSSSIM-VAE [27]	1.22541 ± 0.18918	9.17 ± 0.73	0.167 ± 0.027
VAEGAN [22]	0.86575 ± 0.03080	10.63 ± 0.15	0.073 ± 0.014

images of the CN subjects. Results are reported in Table 1. We observe that no model clearly outperforms the others. On the other hand, VAMP [30], VAE-lin-NF [25], MSSSIM-VAE [27] and VAEGAN [22] perform less well than the others (MSE > 0.05, PSNR < 20 dB, SSIM < 0.5). A possible explanation is that the dataset is too small for these models to learn the data distribution.

The other models obtain a similar performance with, on average, an MSE $<$ 0.04, PSNR > 24 dB and SSIM comprised between 0.69 and 0.75. Not surprisingly, the AE leads to a good performance for this reconstruction task according to the MSE as it is the optimized metric. The vanilla VAE [19] seems to be one of the best models but does not stand out from the other models. It is probable that some models would benefit from hyper-parameter fine tuning to perform better, but it is interesting to see that optimal parameters obtained on classic computer vision datasets do generalize to this different application for many models.

3.2 Pseudo-healthy reconstruction from images simulating dementia

In the following, we discarded the four models that did not give acceptable reconstructions. We first report, for the five dementia subtypes considered simulated with a hypometabolism of 30%, the MSE and SSIM between the simulated image and their reconstructions within the binary mask where hypometabolism was applied (e.g. between X' and $\widehat{X'}$ within the binarized mask M in Fig. 1). All the models reach a very similar performance with an MSE on average across models of 0.0132 (min MSE of 0.0096 for the RAE GP [11] and max MSE of 0.0183 for the IWAE [2]) and an average SSIM of 0.710 (min SSIM of 0.684 for

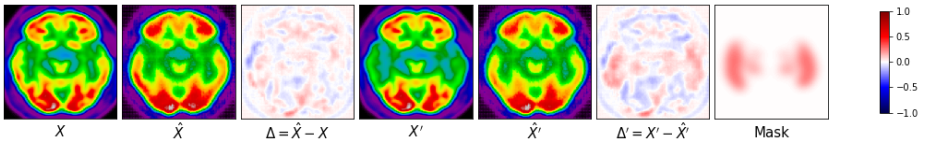

Fig. 1. Example of FDG PET image of a CN subject (X) with the corresponding pseudo-healthy reconstruction ($\widehat{X}$) and difference image (Δ), followed by an image simulating AD hypometabolism obtained from X (X') with the corresponding pseudo-healthy reconstruction ($\widehat{X'}$) and difference image (Δ'), and the mask used to generate X' (M). The pseudo-healthy reconstructions were obtained from the vanilla VAE model.

the IWAE [2] and max SSIM of 0.733 for the RAE-ℓ^2 [11]). This means that the VAE-based models can generalize to various kinds of anomalies located in different parts of the brain, and that none of the tested models can be selected based on this criteria. The average MSE over all the models and all the dementia subtypes (between X' and $\widehat{X'}$) is 0.0132 in the pathological masks M against 0.0072 outside the masks, which makes a 58.6% difference between both regions. The average SSIM is 0.710 inside masks M against 0.772 outside the masks for a 8.4% difference. This shows that the reconstruction error is much larger in regions that have been used for hypometabolism simulation, as expected. For comparison, the percentage difference is only 10.2% for the MSE and 0.2% for the SSIM when computed between the pseudo-healthy reconstruction $\widehat{X'}$ and the real pathology-free images X. This illustrates that the models are all capable of reconstructing the pathological regions as healthy.

We then report in Fig.2 the MSE within the mask simulating AD when generating hypometabolism of various degrees (5% to 70%) for each model. It is interesting to observe that most of the models could be used to detect anomalies of higher intensity as they have an increasing difference in terms of MSE for hypometabolism of 20% and more. The same trend was observed with the SSIM. The RAE-ℓ^2 [11] does not scale as well as other models, probably because the regularization is done on the decoder weights so nothing prevents the encoder from learning a posterior that is less general. We also notice that the IWAE [2] has a worse reconstruction on the pathological region compared to other models, and this becomes more pronounced when the severity of the disease is increased. However this does not mean that IWAE [2] better detects pathological areas since the reconstruction is poor in the whole image as well, meaning that IWAE [2] cannot perform well when the image is out of the training distribution. Surprisingly, the simple autoencoder gives similar results as other methods.

4 Conclusion

The proposed benchmark aimed to introduce the use of recent VAE variants with medical imaging data of high dimension and compare their performance on the detection of dementia-related anomalies on 3D FDG PET brain images. We

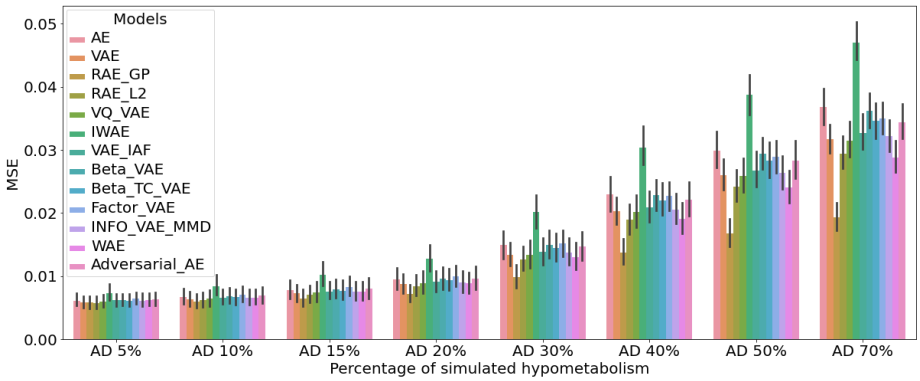

Fig. 2. Bar plot of the evolution of the MSE when computed within the mask characteristic of AD between the image simulated with different degrees of hypometabolism and its reconstruction. We observe that most models can scale to large anomalies.

observed that most models have a comparable reconstruction ability when fed with images of healthy subjects and that their outputs correspond to healthy looking images when fed with images simulating anomalies. Exceptions are the VAEGAN [22], VAMP [30], VAE-lin-NF [25], MSSSIM-VAE [27], RAE-ℓ^2 [11] and IWAE [2]. Thanks to the evaluation framework that consists in simulating images with anomalies from pathology-free images, we showed that most models can generalize pseudo-healthy reconstruction to different dementias and different severity degrees. These results are interesting as it means that VAE-based models developed for natural images can generalize well to other tasks (here 3D brain imaging): they are easy to use and do not necessarily require a large training set, which might not be the case for other types of generative models. We also showed that in our scenario (small dataset of complex 3D images) the simplest models (vanilla AE and VAE) lead to results comparable to that of the more complex ones. Nevertheless, the results are for now limited to the detection of simulated anomalies. An evaluation on real images would be necessary to confirm these observations.

The proposed benchmark could be used in future work to assess whether the posterior learned by the different models is the same for images from healthy and diseased subjects using the simulation framework to compare the latent representation of both the original and simulated images, thus explaining the results of the models. It would also be interesting to compare some of the VAE-based models to GANs or diffusion models, and assess whether it would be possible to improve reconstruction quality while learning the distribution of healthy subject images.

5 Acknowledgment

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). This work was granted access to the HPC resources of IDRIS under the allocation AD011011648 made by GENCI. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).

References

1. Baur, C., Denner, S., Wiestler, B., Navab, N., Albarqouni, S.: Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study. MedIA **69**, 101952 (2021)
2. Burda, Y., Grosse, R.B., Salakhutdinov, R.: Importance weighted autoencoders. In: ICLR (2016)
3. Burgos, N., Cardoso, M.J., Samper-González, J., Habert, M.O., Durrleman, S., Ourselin, S., Colliot, O.: Anomaly detection for the individual analysis of brain PET images. J Med Imag **8**(2), 024003 (2021)
4. Chadebec, C., Vincent, L.J., Allassonniere, S.: Pythae: Unifying generative autoencoders in python - a benchmarking use case. In: Thirty-sixth Conference on NeurIPS Datasets and Benchmarks Track (2022)
5. Chen, R.T., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. Advances in NeurIPS **31** (2018)
6. Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. In: MIDL (2018)
7. Chen, X., Konukoglu, E.: Unsupervised abnormality detection in medical images with deep generative methods, pp. 303–324. Elsevier (2022)
8. Chételat, G., Arbizu, J., Barthel, H., Garibotto, V., Law, I., Morbelli, S., van de Giessen, E., Agosta, F., Barkhof, F., Brooks, D.J., et al.: Amyloid-PET and 18F-FDG-PET in the diagnostic investigation of Alzheimer's disease and other dementias. The Lancet Neurology **19**(11), 951–962 (2020)
9. Choi, H., Ha, S., Kang, H., Lee, H., Lee, D.S.: Deep learning only by normal brain PET identify unheralded brain anomalies. EBioMedicine **43**, 447–453 (2019)
10. Fernando, T., Gammulle, H., Denman, S., Sridharan, S., Fookes, C.: Deep learning for medical anomaly detection – a survey. ACM Computing Surveys **54**(7) (2021)
11. Ghosh, P., Sajjadi, M.S., Vergari, A., Black, M., Schölkopf, B.: From variational to deterministic autoencoders. arXiv:1903.12436 (2019)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in NeurIPS. vol. 27 (2014)
13. Hassanaly, R., Bottani, S., Sauty, B., Colliot, O., Burgos, N.: Simulation-based evaluation framework for deep learning unsupervised anomaly detection on brain FDG PET. In: SPIE Medical Imaging (2023)
14. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-VAE: Learning basic visual concepts with a constrained variational framework. In: ICLR (2017)

15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in NeurIPS **33**, 6840–6851 (2020)
16. 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., Dubois, B.: A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers. Neurology **87**(5), 539–547 (2016)
17. Jagust, W.J., Bandy, D., Chen, K., Foster, N.L., Landau, S.M., Mathis, C.A., Price, J.C., Reiman, E.M., Skovronsky, D., Koeppe, R.A.: The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core. Alzheimer's & Dementia **6**(3), 221–229 (2010)
18. Kim, H., Mnih, A.: Disentangling by factorising. In: ICML. pp. 2649–2658. PMLR (2018)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR 2014 - arXiv:1312.6114 (2014)
20. Kingma, D.P., Welling, M.: An Introduction to Variational Autoencoders. now publishers Inc (2019)
21. Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M.: Improved variational inference with inverse autoregressive flow. Advances in NeurIPS **29** (2016)
22. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: ICML. pp. 1558–1566. PMLR (2016)
23. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. arXiv:1511.05644 (2015)
24. Mostapha, M., Prieto, J., Murphy, V., Girault, J., Foster, M., Rumple, A., Blocher, J., Lin, W., Elison, J., Gilmore, J., et al.: Semi-supervised vae-gan for out-of-sample detection applied to mri quality control. In: MICCAI. pp. 127–136. Springer (2019)
25. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: ICML. pp. 1530–1538. PMLR (2015)
26. 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., Colliot, O.: Clinica: An open-source software platform for reproducible clinical neuroscience studies. Front Neuroinform **15** (2021)
27. Snell, J., Ridgeway, K., Liao, R., Roads, B.D., Mozer, M.C., Zemel, R.S.: Learning to generate images with perceptual similarity metrics. In: ICIP. pp. 4277–4281. IEEE (2017)
28. Thibeau-Sutre, E., Díaz, M., Hassanaly, R., Routier, A., Dormont, D., Colliot, O., Burgos, N.: Clinicadl: An open-source deep learning software for reproducible neuroimaging processing. Comput Meth Prog Bio **220**, 106818 (2022)
29. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. In: ICLR (2018)
30. Tomczak, J., Welling, M.: VAE with a VampPrior. In: International Conference on Artificial Intelligence and Statistics. pp. 1214–1223. PMLR (2018)
31. Uzunova, H., Schultz, S., Handels, H., Ehrhardt, J.: Unsupervised pathology detection in medical images using conditional variational autoencoders. IJCARS **14**, 451–461 (2019)
32. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in NeurIPS **30** (2017)

33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE T Image Process **13**(4), 600–612 (2004)
34. Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., Dormont, D., Durrleman, S., Burgos, N., Colliot, O.: Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. MedIA **63**, 101694 (2020)
35. Xia, T., Chartsias, A., Tsaftaris, S.A.: Pseudo-healthy synthesis with pathology disentanglement and adversarial learning. MedIA **64**, 101719 (2020)
36. Zhao, S., Song, J., Ermon, S.: Infovae: Balancing learning and inference in variational autoencoders. In: Proc AAAI conference on artificial intelligence. vol. 33, pp. 5885–5892 (2019)

Unsupervised anomaly detection in 3D brain FDG PET: A benchmark of 17 VAE-based approaches

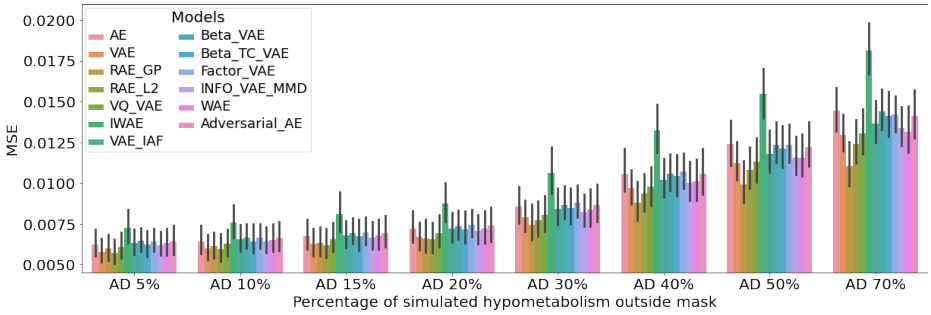

Fig. 1. Evolution of the MSE when computed within the brain but outside the mask characteristic of AD between the image simulated with different degrees of hypometabolism and its reconstruction. We observe that the MSE is not increasing as much as inside the mask characteristic of AD when simulating more severe hypometabolism (Fig. 2).

Table 1. MSE and SSIM inside and outside the mask used for simulation, averaged across all the simulated dementias (30%) for all the models giving acceptable reconstruction.

	Within pathological mask M		Within brain but outside mask M	
Models	MSE	SSIM	MSE	SSIM
AE	0.01385 ± 0.00798	0.714 ± 0.043	0.00727 ± 0.00356	0.775 ± 0.024
VAE [19]	0.01281 ± 0.00724	0.725 ± 0.044	0.00679 ± 0.00307	0.786 ± 0.023
RAE-GP [11]	0.00963 ± 0.00642	0.729 ± 0.046	0.00662 ± 0.00346	0.784 ± 0.025
RAE-ℓ^2 [11]	0.01203 ± 0.00727	0.733 ± 0.043	0.00665 ± 0.00332	0.793 ± 0.024
VQVAE [32]	0.01269 ± 0.00770	0.720 ± 0.047	0.00696 ± 0.00348	0.785 ± 0.025
IWAE [2]	0.01839 ± 0.00929	0.684 ± 0.048	0.00882 ± 0.00445	0.747 ± 0.026
VAE-IAF [21]	0.01294 ± 0.00756	0.705 ± 0.047	0.00725 ± 0.00333	0.768 ± 0.025
β-VAE [14]	0.01365 ± 0.00771	0.705 ± 0.047	0.00740 ± 0.00314	0.769 ± 0.025
β-TC VAE [5]	0.01326 ± 0.00771	0.708 ± 0.048	0.00723 ± 0.00354	0.774 ± 0.025
FactorVAE [18]	0.01420 ± 0.00754	0.689 ± 0.048	0.00751 ± 0.00316	0.750 ± 0.025
InfoVAE [36]	0.01295 ± 0.00751	0.713 ± 0.045	0.00711 ± 0.00322	0.777 ± 0.025
WAE [29]	0.01247 ± 0.00718	0.698 ± 0.045	0.00726 ± 0.00350	0.758 ± 0.025
AAE [23]	0.01369 ± 0.00805	0.703 ± 0.047	0.00743 ± 0.00370	0.765 ± 0.025

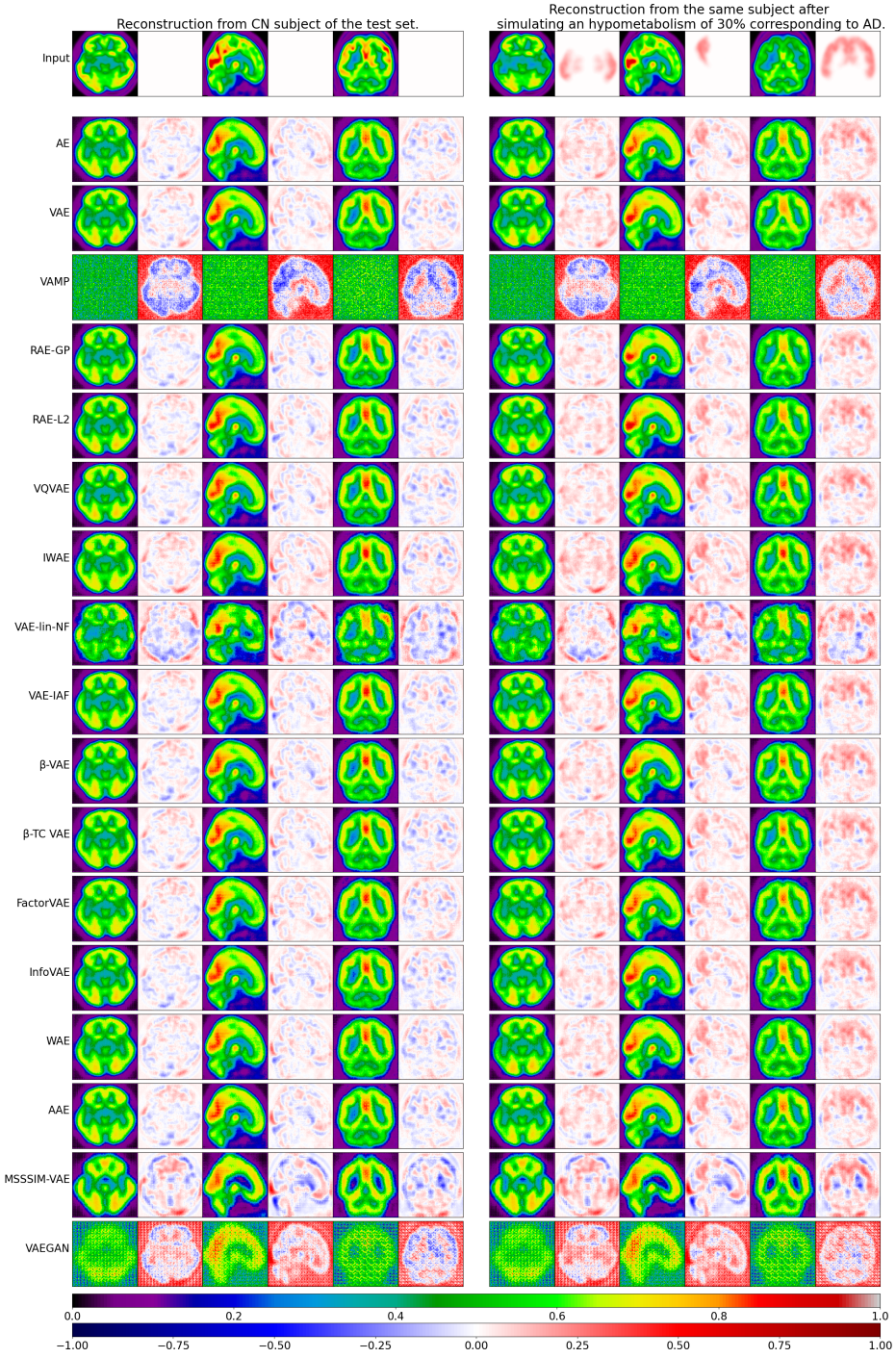

Fig. 2. Reconstruction obtained for each benchmarked model from the real image of a CN subject in the test set (left) and from an image simulating a 30% hypometabolism in the region associated with AD based on the same CN subject (right).