CTFSH: Full Head CT Anomaly Detection with Unsupervised Learning

Michele Guerra[*1] MICHELE.GUERRA@TUM.DE
Abhijeet Parida[*2] ABHIJEET@DEEPC.AI
Daniel Rückert[1] DANIEL.RUEKERT@TUM.DE
[1] *Technical University of Munich, Munich, Germany*
Mehmet Yigitsoy[†2] MEHMET@DEEPC.AI
[2] *deepc GmbH, Munich, Germany*
Shadi Albarqouni[†1,3,4] SHADI.ALBARQOUNI@UKBONN.DE
[3] *Clinic for Diagnostic and Interventional Radiology, University Hospital Bonn, Bonn, Germany*
[4] *Helmholtz AI, Helmholtz Zentrum Muenchen, Neuherberg, Germany*

Editors: Under Review for MIDL 2022

Abstract

Unsupervised Anomaly Detection (UAD) is an inexpensive and effective method to bring value to the clinical workflow for pathology detection, especially in the emergency room setting, where quick prioritization of Computed Tomography (CT) scans is necessary. While there are numerous works dealing with UAD for medical images, most of them focus on Magnetic Resonance Imaging (MRI) and 2D slices of the brain. This work's aim is to build a comparison between two commonly used baselines for UAD of volumetric CT scans. In addition to this, we borrowed two recent contributions to the field of Computer Vision in order to improve the reconstruction quality of our networks. These contributions effectively increased the AUROC for anomaly detection from 0.74 to 0.90 for one of our baselines. In order to guarantee that the anomaly detection algorithm is effective for all pathologies, including fractures, we tested our models both on skull-stripped scans and full-head scans. Leaving the skull in the CT volumes allowed the algorithm to efficiently classify fractures. To the best of our knowledge, this is the first work to show a comparison regarding the usage of skull-stripping. To facilitate further research in UAD for Head CT, we publish supplementary labels for the publicly available CQ500 dataset. The code for this study can be found in a GitHub repository at https://github.com/pederismo/CTFSH

Keywords: Unsupervised Anomaly Detection, One Class Classification, Out of Distribution Detection

1. Introduction

In the past years, there has been a 43% increase in the number of radiographical scans of various modalities (Taschetta-Millane and Fornell, 2021). It is estimated that 10-15% of the scans may be delayed, missed, or incorrectly diagnosed (Bruno et al., 2015), and due to this increased workload, the error rates are expected to grow. The use of Deep Learning based algorithms can help in reducing some of these errors.

[*] Equal Contribution
[†] Senior Authors with Equal Contribution

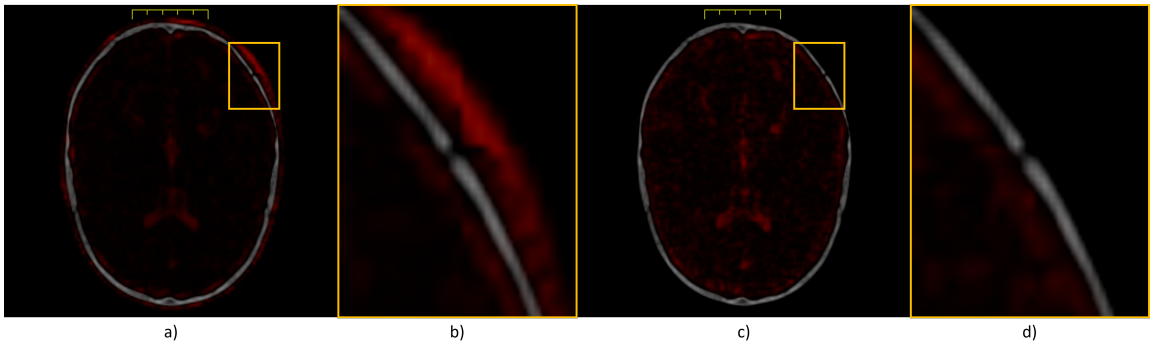

Figure 1: The same patient was analyzed by two versions of one of our baselines. The skull-stripped version correctly classifies the patient as unhealthy. We overlay the heatmaps from the models on top of the skulls. (a) shows the heatmap from the model trained on the full head; (b) zooms-in on the fracture and heatmap from the full head model; (c) shows the heatmap from the model trained on skull-stripped brains; (d) zooms-in on the fracture and heatmap from the skull-stripped model.

These algorithms usually require large-scale fully-annotated datasets. Annotations in the medical domain are performed by highly specialized experts, who have developed their skills through years of practice. Nevertheless, the quality of the annotations may not be optimal due to inter/intra-observer variability (Sampat et al., 2006). Supervised algorithms also require the training and testing distributions to be identical: when presented with a new testing class they haven't seen before, they tend to fail silently by giving a high-confidence prediction (Hendrycks and Gimpel, 2017). Unsupervised Anomaly Detection (UAD) is the best strategy to combat all these learning problems as well as provide useful support to medical professionals to do faster and accurate diagnoses (Bruno et al., 2015; Taschetta-Millane and Fornell, 2021).

So far most studies on AI-assisted diagnosis for the head have been dealing exclusively with MRI scans (Baur et al., 2021) for their high level of detail when visualizing soft tissues. Nevertheless, non-contrast head CT scans are the most commonly used tool in the emergency room for patients with head injury (Chilamkurthy et al., 2018) and some clinical conditions require fast evaluation and treatment (e.g. Intracerebral Hemorrhages (ICH) (Elliott and Smith, 2010), strokes or changes in intracranial pressure (Chilamkurthy et al., 2018)). Therefore, diagnosis of head CT scans is critical and time-sensitive. This calls for a fast and efficient AI tool adapted to CTs that could be integrated into the healthcare workflow.

A common strategy adopted by the research is to work on skull-stripped scans instead of full head scans (Woods et al., 1998). This process may prevent the algorithms from successfully detecting anomalies that may be present in the skull, which is crucial for trauma patients. As seen in Figure 1, models trained on skull-stripped scans can miss detecting fractures. Therefore, it is important to do a full head analysis of the CT scans.

Contributions The main contributions of the paper are that it provides: 1) an efficient methodology for training and evaluating an end-to-end deep learning-based UAD algorithm for the 3D volumetric CT scans of the human brain; 2) the inclusion of two Computer Vision contributions, namely the MS-SSIM (Wang et al., 2003) and ACAI (Berthelot et al., 2019), to the baseline architectures involved in the study, to improve the reconstruction quality of our baselines and to provide smoothness to the underlying manifold; 3) analysis on the impact of skull stripping vs. full head towards anomaly detection in the 3D volumetric CT scans; 4) an extension of the labels of the CQ500 dataset (Chilamkurthy et al., 2018) towards benchmarking of anomaly detection models.

2. Related Works

Most Deep Learning-based approaches for UAD are based on generative models, like the Autoencoder (AE) (Hinton and Salakhutdinov, 2006). The dimensionality reduction provided by the bottleneck of such networks enables generalization across newly, unseen inputs. After successful training of the neural network on normal inputs, the model will fail to reconstruct anomalies during inference. The difference between reconstruction and input will highlight said anomalies (Baur et al., 2021).

VAE While the original AE is good at reconstructing inputs, its simplicity can fail at generalization. Improvements like the Variational Autoencoder (VAE) (Kingma and Welling, 2014) have been made, which uses the KL divergence to enforce similarity between the input distribution and the normal. For anomaly detection, Guo et al. (2021) adopt a cascade-like architecture of VAEs to combine latent representation at multiple scales to achieve better reconstruction quality. Pinaya et al. (2021) instead uses a VQ-VAE and then an attention mechanism with transformers to learn the probability density function of the latent representation of data.

GAN Generative Adversarial Networks (Goodfellow et al., 2014) are another type of generative model that has been adopted for UAD. GANs have been chosen for anomaly detection thanks to their high-quality reconstructions: in clinical optical coherence tomography analysis, Schlegl et al. (2019) first proposed AnoGAN, with a newly designed anomaly score that uses information from the image space and the latent space, and then improved on this work by substituting the Deep Convolutional GAN architecture with a Wasserstein GAN (Arjovsky et al., 2017), and also changed the slow iterative mapping from image to feature space with an encoder. Simarro Viana et al. (2021) adapted this last method to Computed Tomography and 3D for anomaly detection.

MS-SSIM and ACAI The L1 or Mean Absolute Error (MAE) loss is the simplest choice of loss for the training of the AE. Although easy to implement, an AE trained exclusively with the MAE tends to produce blurry reconstructions and does not generalize well. To improve reconstruction quality, Zhao et al. (2017) introduced a combination of the L1-loss and the Multi-Scale Structural Similarity Index (MS-SSIM) (Wang et al., 2003). The MS-SSIM as a loss function is sensitive to changes in local structures, while being differentiable, and therefore appropriate for backpropagation.

The MS-SSIM has already been investigated for Unsupervised Defect Segmentation (Bergmann. et al., 2019), but so far nobody has used it for medical imaging. Although the

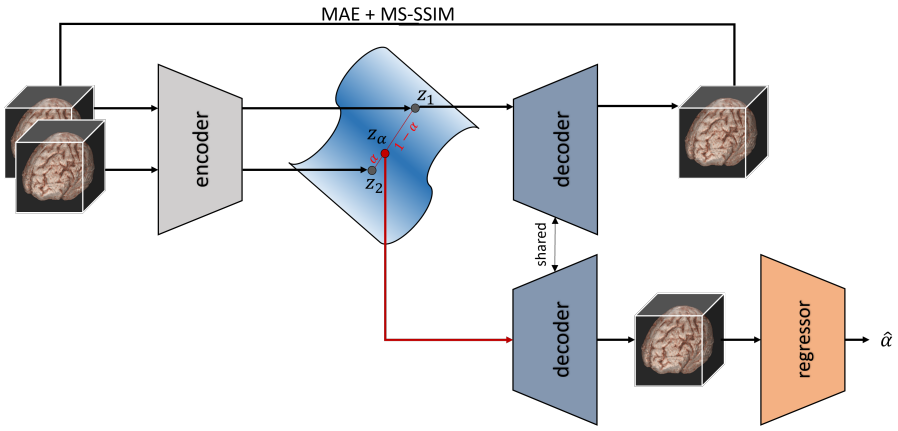

Figure 2: The AE model after the addition of MS-SSIM and ACAI at training. Two incoming volumes are encoded and then interpolated. After the decoder has reconstructed the latent interpolation, the critic is trained to regress the α used to interpolate. Training of the AE happens thanks to the MAE and MS-SSIM loss.

MS-SSIM helps achieve visually pleasing reconstructions, it does not ensure the quality of the latent manifold. Berthelot et al. (2019) proposed an Adversarially Constrained Autoencoder Interpolation (ACAI) that improves generalization and quality of the latent space by enabling the AE to interpolate better between two points in the manifold (see Fig. 2). The ability to create high quality interpolations that smoothly transition from point z_1 to point z_2, demonstrates that the network has gained knowledge of the manifold's structure.

Anomaly Score Once the model is trained, it is tested on unseen data, both healthy and anomalous, to see if it can correctly classify the input volumes. Calculating a single score to discriminate between healthy and unhealthy samples is not an easy task. Most studies that rely on generative models have adopted the L1 or L2 norm between input and reconstruction as a score, similarly to the loss function used in training (Schlegl et al., 2019; Simarro Viana et al., 2021). Pinaya et al. (2021), on the other hand, relies on the likelihood scores generated by the transformer and averages them. Ruff et al. (2018) and Chen et al. (2021) proposed a way of calculating anomaly scores from the latent space codes, by computing how distant they are from a hypersphere center or from a Gaussian distribution, respectively.

3. Methodology

Given a dataset $\mathcal{D} = \mathcal{H} \cup \mathcal{A}$, which is the union of healthy set $\mathcal{H}$ and anomalous set $\mathcal{A}$, such that $\mathcal{H} \cap \mathcal{A} = \varnothing$, we train a generative model composed of an encoder $f_\phi(\cdot)$ and a decoder $g_\theta(\cdot)$ on $\mathcal{H} = \{x_i\}_{i=1}^{|\mathcal{H}|}$, where $x_i \in \mathbb{R}^{H \times W \times C}$, and each x_i represents the i^{th} 3D healthy brain volume of spatial resolution $(H \times W)$ and C slices. The implementations for the encoder and the decoder are described in Appendix D.

The generative models studied are two: an AE where $\hat{x} = g_\theta(f_\phi(x))$ and a VAE where $\mu_{VAE}, \sigma_{VAE} = f_\phi(x)$, $z \sim \mathcal{N}(\mu_{VAE}, \sigma_{VAE}$, $\hat{x} = g_\theta(z)$. Two distinct samples are drawn from the dataset $\mathcal{H}$, $x_1, x_2 \sim \mathcal{H}$. The encoder $f_\phi(\cdot)$ is used to extract the latent codes $z_1 = f_\phi(x_1)$ and $z_2 = f_\phi(x_2)$, where $z_1, z_2 \in \mathbb{R}^d$. By linear interpolation of z_1 and z_2, a new code $z_\alpha = \alpha \cdot z_1 + (1-\alpha) \cdot z_2$, with $\alpha \sim \mathcal{U}_{[0,0.5]}$ is obtained, as it can be seen from Figure 2. The decoder $g_\theta(\cdot)$ now takes the latent codes z_1 and z_α as input and reconstructs the full 3D healthy scans $\hat{x}_1, \hat{x}_\alpha \in \mathbb{R}^{H \times W \times C}$, such that $\hat{x}_1 = g_\theta(z_1)$ and $\hat{x}_\alpha = g_\theta(\hat{z}_\alpha)$. A regressor $c_\psi(\cdot)$ tries to predict the α used for the interpolation, $\hat{\alpha} = c_\psi(\hat{x}_\alpha)$, with $\hat{\alpha} \in \mathbb{R}$. The action of the regressor helps to get a smooth manifold, where the latent representation $z_i \in \mathbb{R}^d$ of each healthy brain resides. The training step for the encoder and decoder is done by minimizing the following loss borrowed from Zhao et al. (2017):

$$\ell_r(x, \hat{x}) = \rho \cdot \ell_{MAE}(x, \hat{x}) + (1-\rho) \cdot \ell_{MS\text{-}SSIM}(x, \hat{x}), \tag{1}$$

where $\rho = 0.15$. The regressor, on the other hand, is trained on the ℓ_2-norm between the sampled α and the predicted one:

$$\ell_d(\alpha, \hat{\alpha}) = \|\alpha - \hat{\alpha}\|_2^2 . \tag{2}$$

During inference, the trained encoder $f_\phi(\cdot)$ transforms an unhealthy sample $x_1' \sim \mathcal{A}$ to the nearest healthy latent code, $z_1 = f_\phi(x_1')$. The reconstruction $\hat{x}_1 = g_\theta(z_1)$ is therefore a healthy reconstruction of the unhealthy input x_1', and this difference shows when calculating $\ell_r(x_1', \hat{x}_1)$. On the other hand, for a healthy sample $x_1 \sim \mathcal{H}$, the reconstruction $\hat{x}_1 = g_\theta(f_\phi(x_1))$ will be closer to x_1, thus producing a lower $\ell_r(x_1, \hat{x}_1)$. Therefore we adopted Equation 1 both as the training loss and as anomaly score, which we call **I** score.

Each method was also designed to generate a prediction based on the latent codes. This **F** score, inspired by the work of Chen et al. (2021), was used at inference to measure the distance between the testing latent codes and a Gaussian distribution built upon the manifold of the training healthy samples:

$$\mathbf{F}_i = 1 - exp(-\frac{\|f_\phi(x_i) - \boldsymbol{\mu}\|_2^2}{\sigma^2}), \tag{3}$$

with $\mu = \frac{1}{|\mathcal{H}|} \sum_{i=1}^{|\mathcal{H}|} f_\phi(x_i) \in \mathbb{R}^d$ and $\sigma^2 = \frac{1}{|\mathcal{H}|} \sum_{i=1}^{|\mathcal{H}|} \|f_\phi(x_i) - \mu\|_2^2 \in \mathbb{R}$ calculated on the healthy samples.

4. Experiments and Results

Dataset The CQ500 dataset (Chilamkurthy et al., 2018) is used for training and validation. It contains 491 CT scans of the human brain collected at different radiology centers in New Delhi, using different scanners. It was originally designed to be a validation dataset for the detection of ICH, midline shift, mass effect or calvarial fractures. In order to be used for the development of a more comprehensive detection algorithm, the dataset was relabeled with the addition of pathologies like fractures, ischemia and atrophies. The new extra labels were obtained following a majority vote by three expert radiologists. The details on the anomalous samples can be found in Appendix A.

Experimental Setup The healthy subset $\mathcal{H}$ of the dataset was split in three parts for training (80%), validation (10%), and testing (10%) of our methods. The preprocessing pipeline and the implementation details are reported in Appendix B and D. The unhealthy subset $\mathcal{A}$ was further divided into each pathology subset, such that $\mathcal{A} = \mathcal{A}_{ATY} \cup \mathcal{A}_{FRAC} \cup \mathcal{A}_{ICH} \cup \mathcal{A}_{ISCH} \cup \mathcal{A}_{MASS} \cup \mathcal{A}_{OTHER}$. For each unhealthy subset $\mathcal{A}_i$ of $\mathcal{A}$, the model was tested against the testing healthy dataset ($n = 12$) and $\mathcal{A}_i$. For each scan the models computed two predictions, the **I** score, defined in Equation 1, and the **F** score, defined in Equation 3. Both scores were then used independently to compute the Area Under the Receiver Operating Characteristics curve (AUROC(I) and AUROC(F)).

Table 1: The comparison of AUROC scores between the AE and the VAE on different pathologies.

Pathology (n. scans)	**Method**	***L1***		***L1 + MS-SSIM***		***L1 + MS-SSIM + ACAI***	
		AUROC(I)	AUROC(F)	AUROC(I)	AUROC(F)	AUROC(I)	AUROC(F)
Atrophy	AE	0.739	0.361	0.894	0.372	**0.900**	0.439
(15)	VAE	0.722	0.517	0.883	0.483	0.894	0.458
Fracture	AE	**0.744**	0.410	0.724	0.321	0.689	0.522
(26)	VAE	0.734	0.465	0.712	0.365	0.596	0.479
ICH	AE	0.708	0.386	0.779	0.370	**0.784**	0.496
(155)	VAE	0.670	0.441	0.768	0.427	0.731	0.509
Ischemia	AE	0.812	0.461	0.857	0.446	**0.881**	0.576
(59)	VAE	0.778	0.533	0.843	0.533	0.822	0.511
Mass	AE	0.492	0.450	0.575	0.558	0.542	**0.633**
(10)	VAE	0.500	0.458	0.592	0.558	0.525	0.513
Other	AE	0.639	0.403	0.642	0.397	0.642	0.522
(30)	VAE	0.622	0.450	**0.650**	0.456	0.628	0.476
TOTAL	AE	0.703	0.399	0.765	0.389	**0.771**	0.515
	VAE	0.673	0.461	0.756	0.451	0.724	0.506

4.1. Effects of MS-SSIM and ACAI

By adding the MS-SSIM and ACAI the results consistently improve on the **I** scores for all pathologies except Fracture (see Tables 1, 2, 3. In Figure 3 we can look at the reconstruction quality for three methods compared (complete comparison can be found in Appendix E). It can be seen how the MS-SSIM delivers on its purpose of generating visually pleasing reconstructions that improve the **I** scores. Additional motivation behind the choice of including MS-SSIM is brought by Figure 4, where it can be seen how the **I** scores, after addition of MS-SSIM, allow for easier discrimination between healthy and unhealthy data by increasing the gap between the mean of healthy **I** scores, and the mean of all other classes.

4.2. AE vs VAE

Table 1 shows that for unstripped test data the AE consistently outperforms the VAE all pathologies except Mass Effect (L1 and L1+MS-SSIM) and Other (L1+MS-SSIM). This can be attributed to improper balancing between the reconstruction error and the KL divergence. Reconstruction quality can be improved by finding the best weight between

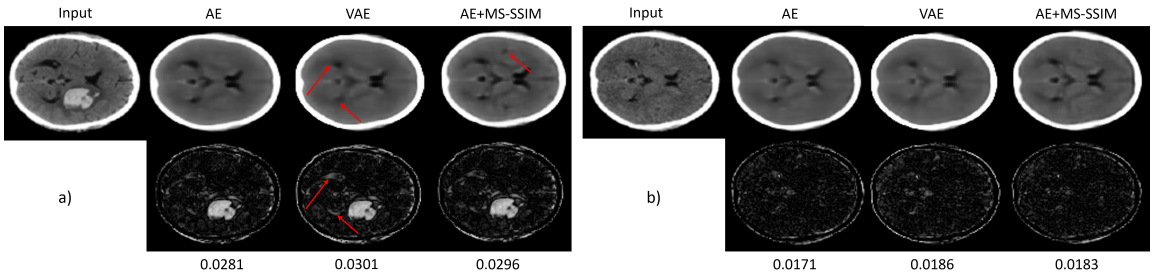

Figure 3: Comparison for reconstruction quality on: a) a patient with ICH; b) an healthy patient. Brightness has been increased on the residual maps to better visualize the differences. Pointed is an example of how the VAE fails to reconstruct the occipital horns of the lateral ventricles. Below are reported the L1 values for each reconstruction.

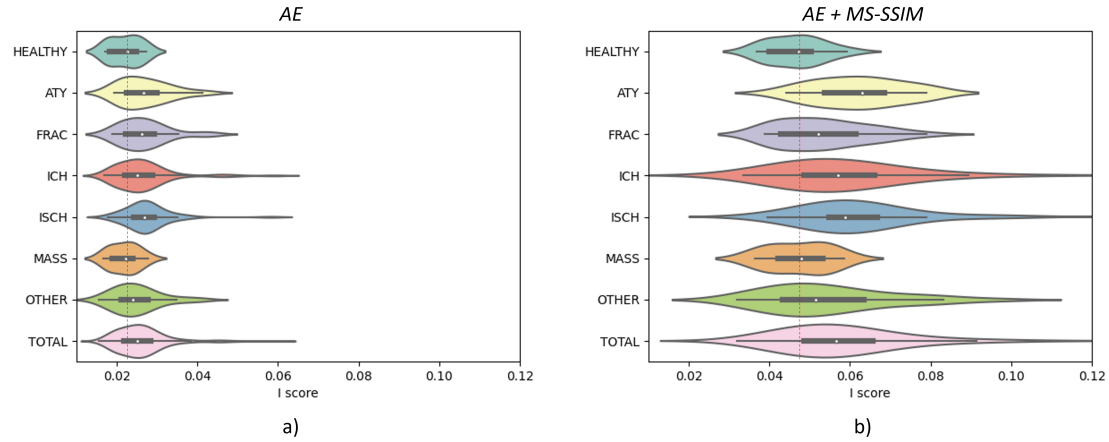

Figure 4: A boxplot and a kernel density estimation for the **I** scores of two AE versions. For each testing class we saved the predictions and plotted them. For Atrophy and Ischemia the gap that separates them from Healthy is bigger.

these loss components (Dai and Wipf, 2018; Asperti and Trentin). In addition to that, all models struggle to detect anomalies of the Mass Effect category (e.g. tumors, cavernoma). A reason could be that tumors are hypodense anomalies where the voxel value is darker than its surrounding, resulting in a reduced contrast in the residual error. Further study on how to correctly detect such anomalies on CT scans has to be carried out.

4.3. Effects of Skull Stripping

Looking at Table 2, as expected, the performance on Fracture detection increases when we train and test our baselines on full-head data. To visualize this result, we took the same

scan, both full-head and skull-stripped, and used the AE and the StripAE respectively to generate two heatmaps. These were overlaid on top of the full-head scan. From Figure 1.b it can be seen how the full-head model highlights anomalies related to the fracture, whereas the skull-stripped one does not (1.d). However, the full-head model still shows improvement in the AUROC scores for pathologies other than Fractures, like Mass Effect and Intracranial Hemorrhages, as seen in Table 2.

Table 2: The comparison of AUROC scores between an AE and the same AE trained on skull-stripped data for different pathologies.

Pathology (n. scans)	**Method**	***L1***		***L1+MS-SSIM***		***L1+MS-SSIM+ACAI***	
		AUROC(I)	AUROC(F)	AUROC(I)	AUROC(F)	AUROC(I)	AUROC(F)
Atrophy	AE	0.739	0.361	0.894	0.372	0.900	0.439
(15)	StripAE	0.844	0.556	0.944	0.683	**0.944**	0.750
Fracture	AE	**0.744**	0.410	0.724	0.321	0.689	0.522
(26)	StripAE	0.484	0.381	0.561	0.391	0.522	0.436
ICH	AE	0.708	0.386	0.779	0.370	**0.784**	0.496
(155)	StripAE	0.638	0.398	0.774	0.391	0.751	0.465
Ischemia	AE	0.812	0.461	0.857	0.446	0.881	0.576
(59)	StripAE	0.809	0.551	0.883	0.622	**0.886**	0.679
Mass	AE	0.492	0.450	0.575	0.558	0.542	**0.633**
(10)	StripAE	0.542	0.558	0.617	0.500	0.542	0.533
Other	AE	0.639	0.403	0.642	0.397	0.642	0.522
(30)	StripAE	0.583	0.511	**0.656**	0.589	0.625	0.642
TOTAL	AE	0.703	0.399	0.765	0.389	0.771	0.515
	StripAE	0.662	0.449	**0.774**	0.465	0.756	0.529

4.4. Detection based on Latent Codes

From Tables 1 and 2 the **I** scores outperform the **F** except for *Mass Effect*. However, if we train a VAE, or an AE with ACAI, we have consistent improvements in the **F** scores performance with respect to the vanilla AE. This observation highlights that the detection based on latent space improves if the method is trained on a loss component that focuses on latent representations. Future works should include a new loss function designed on latent representations, in order to harness the full capability of the **F** score.

5. Conclusion

This work aims to identify and explain the behaviour of AutoEncoders and their variations for Unsupervised Anomaly Detection. We tackled full head volumetric CT scans, and gave insights on the effect of skull-stripping. We demonstrated how a model trained of full head scans can outperform the same architecture trained on the skull-stripped scans for anomaly detection of fractures. We showed how simple modifications to the anomaly score estimation and the training losses can affect the detection performance, especially how the MS-SSIM can improve the reconstruction quality and help discriminate better between healthy and unhealthy scans. In the hope of enabling future studies in anomaly detection on the CQ500 dataset, we publish a finer set of labels for the already available public dataset.

Acknowledgments

The authors would like to acknowledge the insights provided by Dr. Gabriel Varjão Lima, MD, as a radiologist into the outputs of the various trained anomaly detection method. Further, the authors are thankful to deepc GmbH for providing the computing resources and sharing the detailed annotations on the CQ500 dataset for the project.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223. PMLR, 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html. ISSN: 2640-3498.

Andrea Asperti and Matteo Trentin. Balancing reconstruction error and kullback-leibler divergence in variational autoencoders. 8:199440–199448. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.3034828. Conference Name: IEEE Access.

Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study. *Medical Image Analysis*, 69:101952, 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101952. URL https://www.sciencedirect.com/science/article/pii/S1361841520303169.

Paul Bergmann., Sindy Löwe., Michael Fauser., David Sattlegger., and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,*, pages 372–380. INSTICC, SciTePress, 2019. ISBN 978-989-758-354-4. doi: 10.5220/0007364503720380.

David Berthelot, Colin Raffel, Aurko Roy, and Ian J. Goodfellow. Understanding and improving interpolation in autoencoders via an adversarial regularizer. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=S1fQSiCcYm.

M. A. Bruno, E. A. Walker, and H. H. Abujudeh. Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. *Radiographics*, 35(6):1668–1676, 10 2015.

Yuanhong Chen, Yu Tian, Guansong Pang, and Gustavo Carneiro. Deep one-class classification via interpolated gaussian descriptor, 2021.

Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G. Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet*, 392(10162):2388–2396, 2018. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(18)31645-3. URL https://www.thelancet.com/journals/

lancet/article/PIIS0140-6736(18)31645-3/fulltext#articleInformation. Publisher: Elsevier.

Bin Dai and David Wipf. Diagnosing and enhancing VAE models. 2018. URL https://openreview.net/forum?id=B1e0X3C9tQ.

Justine Elliott and Martin Smith. The acute management of intracerebral hemorrhage: A clinical review. *Anesthesia & Analgesia*, 110(5):1419–1427, 2010. ISSN 0003-2999. doi: 10.1213/ANE.0b013e3181d568c8. URL https://journals.lww.com/anesthesia-analgesia/Fulltext/2010/05000/The_Acute_Management_of_Intracerebral_Hemorrhage_.28.aspx.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Xiaoyuan Guo, Judy Wawira Gichoya, Saptarshi Purkayastha, and Imon Banerjee. CVAD: A generic medical anomaly detector based on cascade VAE. *arXiv:2110.15811 [cs, eess]*, 2021. URL http://arxiv.org/abs/2110.15811.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1127647. URL https://www.science.org/doi/10.1126/science.1127647.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Mohamed Najm, Hulin Kuang, Alyssa Federico, Uzair Jogiat, Mayank Goyal, Michael D. Hill, Andrew Demchuk, Bijoy K. Menon, and Wu Qiu. Automated brain extraction from head CT and CTA images using convex optimization with shape propagation. *Computer Methods and Programs in Biomedicine*, 176:1–8, 2019. ISSN 0169-2607. doi: 10.1016/j.cmpb.2019.04.030. URL https://www.sciencedirect.com/science/article/pii/S016926071930375X.

Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. 1(10):e3, 2016. ISSN 2476-0757. doi: 10.23915/distill.00003. URL http://distill.pub/2016/deconv-checkerboard.

Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sébastien Ourselin, and M. Jorge Cardoso. Unsupervised brain anomaly detection and segmentation with transformers. In *Proceedings of the Fourth Confer-*

ence on Medical Imaging with Deep Learning, pages 596–617. PMLR, 2021. URL https://proceedings.mlr.press/v143/pinaya21a.html. ISSN: 2640-3498.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4393–4402. PMLR, 2018. URL https://proceedings.mlr.press/v80/ruff18a.html. ISSN: 2640-3498.

Mehul P Sampat, Zhou Wang, Mia K Markey, Gary J Whitman, Tanya W Stephens, and Alan C Bovik. Measuring intra-and inter-observer agreement in identifying and localizing structures in medical images. In *2006 International Conference on Image Processing*, pages 81–84. IEEE, 2006.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.01.010. URL https://www.sciencedirect.com/science/article/pii/S1361841518302640.

Jaime Simarro Viana, Ezequiel de la Rosa, Thijs Vande Vyvere, David Robben, Diana M. Sima, and CENTER-TBI Participants and Investigators. Unsupervised 3d brain anomaly detection. In Alessandro Crimi and Spyridon Bakas, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Lecture Notes in Computer Science, pages 133–142. Springer International Publishing, 2021. ISBN 978-3-030-72084-1. doi: 10.1007/978-3-030-72084-1_13.

M. Taschetta-Millane and D. Fornell. Top trend takeaways in radiology from rsna 2020, 2021. URL http://archive.today/DoawQ.

Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003. doi: 10.1109/ACSSC.2003.1292216.

Roger P. Woods, Scott T. Grafton, John D. G. Watson, Nancy L. Sicotte, and John C. Mazziotta. Automated image registration: Ii. intersubject validation of linear and nonlinear models. *Journal of Computer Assisted Tomography*, 22(1), 1998. ISSN 0363-8715. URL https://journals.lww.com/jcat/Fulltext/1998/01000/Automated_Image_Registration__II__Intersubject.28.aspx.

Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017. ISSN 2333-9403. doi: 10.1109/TCI.2016.2644865. Conference Name: IEEE Transactions on Computational Imaging.

Appendix A. Extended Annotation of CQ500

The new labels of the CQ500 dataset can be found in *.csv* format in the paper's repository at https://github.com/pederismo/CTFSH. The main focus of the relabeling were pathologies like *Atrophy*, *Bleeding*, *Fracture*, *Ischemia* and *Mass Effect*. Another category called *Others* is added to include patients with anomalies that are not covered by the aforementioned pathologies. Some patients may have more than one conditions and have therefore been included in multiple categories. None of the healthy samples belong to any of the previously mentioned classes.

Appendix B. Preprocessing

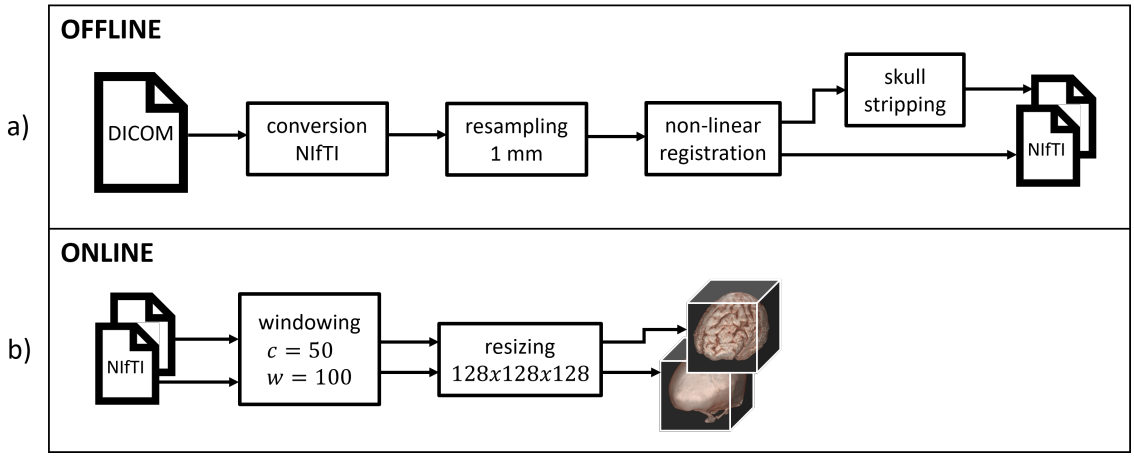

Figure 5: The full preprocessing pipeline adopted for each scan. a) The DICOM file was offline converted to NIfTI, before being registered, resampled and stripped (if necessary); b) before being fed to the network, each scan was additionally windowed and resized, to fit the memory requirements.

The CQ500 scans required a few preprocessing steps to allow for the method to work. The first half of preprocessing was done offline (see Figure 5.a). First of all, the DICOM files were converted to NIfTI. Once this operation is done, the scan was first isotropically resampled to 1 mm voxels, and then registered using a non-linear registration algorithm. Before saving the scans into NIfTI files, each one was either stripped of the skull or left unstripped. The stripping was performed with MATLAB using the method provided by (Najm et al., 2019). The second half of preprocessing steps is done online, while training or testing the algorithms (Figure 5.b). Each scan, whether left with the skull or stripped of it, is windowed using level $c = 50$ and width $w = 100$, before voxel values are normalized between 0 and 1. To enable training in a 3D setting, the input scans had to be resized to satisfy the memory requirements. The size chosen was $128 \times 128 \times 128$, which still allows for good detail quality.

Appendix C. Interpolations

The interpolation between two distinct latent codes shows that there is a smooth manifold onto which the codes are mapped to. We observed no significant difference in the manifold interpolation between the AE with and without the addition of the ACAI, as seen in Figure 6. This could be because AEs may already show signs of interpolation according to Berthelot et al. (2019). Therefore the effect of the ACAI is negligible. In further studies, a new distance metrics in the image space should be defined to make sure the interpolations only happens between input that are close together in the manifold.

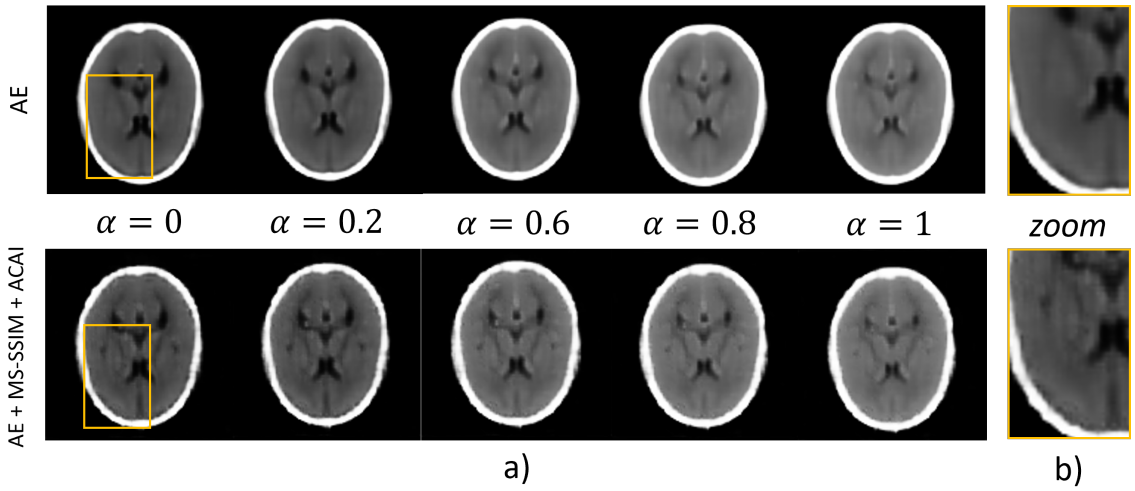

Figure 6: The effect of interpolation on the latent codes. (a) shows different degrees of interpolation between the latent vectors of two different images; (b) zooms-in on regions of the reconstructions that are more visually pleasant thanks to the contribution from MS-SSIM and ACAI.

Appendix D. Implementation Details

All convolution and upsampling operations are adapted to deal with 3D inputs. The encoder is made of five convolution operations with kernel size $k_e = (4, 4, 4)$, stride $s_e = 2$, dilation $d_e = 2$ and respectively $(32, 64, 128, 256, 512)$ channels. For an input size of $1 \times 128 \times 128 \times 128$, the encoder therefore ends up generating a $512 \times 4 \times 4 \times 4$ latent representation which then gets flattened before being fed to a last fully connected layer that reduces the latent size to $l = 512$. For the decoder, parameters were chosen accordingly in order to reduce checkerboard artifacts in the reconstructions (Odena et al., 2016). Each transposed convolution was divided into an initial upsampling operation followed by a regular convolution with activation function. Nearest neighbors and a scale factor of 2 was chosen for each upsampling operation. Convolutions have a kernel size $k_d = 3$, stride $s_d = 1$ and dilation $d_d = 1$. The channels mirror the encoder ones. For both sides of the architecture we adopted Leaky ReLU as activation function, and Instance Normalization.

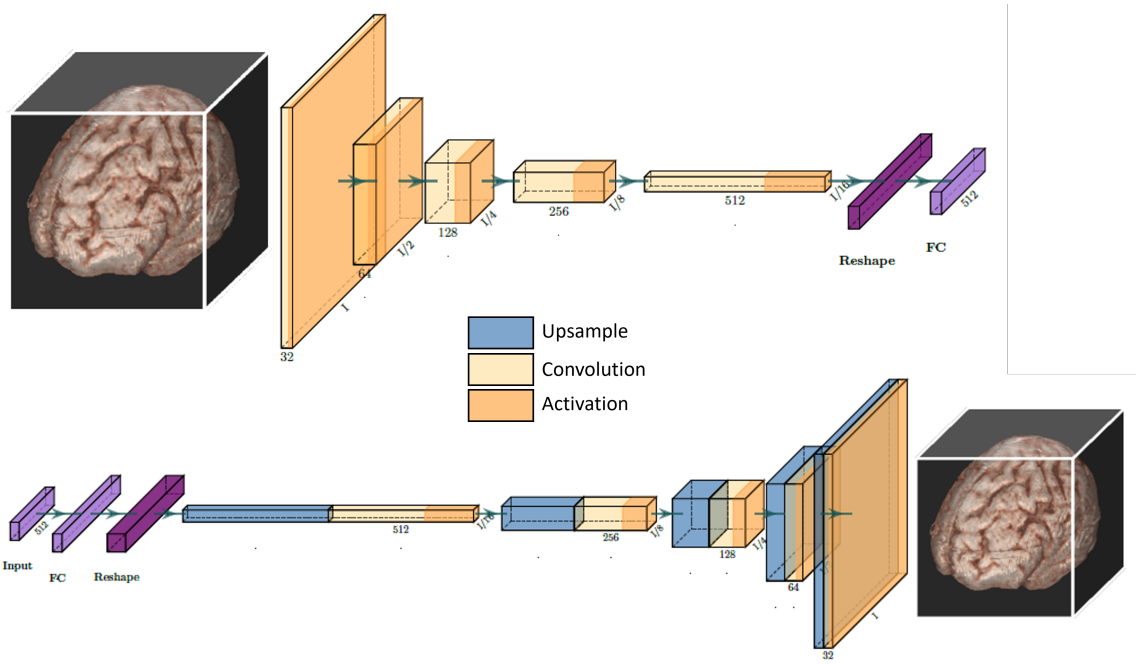

Figure 7: The architecture chosen for all the methods involved in the work. Convolution and upsample operations are for 3D input data.

Each method was trained for 200 epochs on a NVIDIA GeForce RTX 2070 with a learning rate $lr = 0.0001$. For the methods that included the MS-SSIM, the weight used in Equation 1 between the two loss components was $\rho = 0.15$. For the ACAI components, $\lambda = 1.$ was chosen for the AE regularizer and $\zeta = 0.2$ for the critic's one.

Appendix E. Full Reconstruction Quality Comparison

Appendix F. Additional Tables

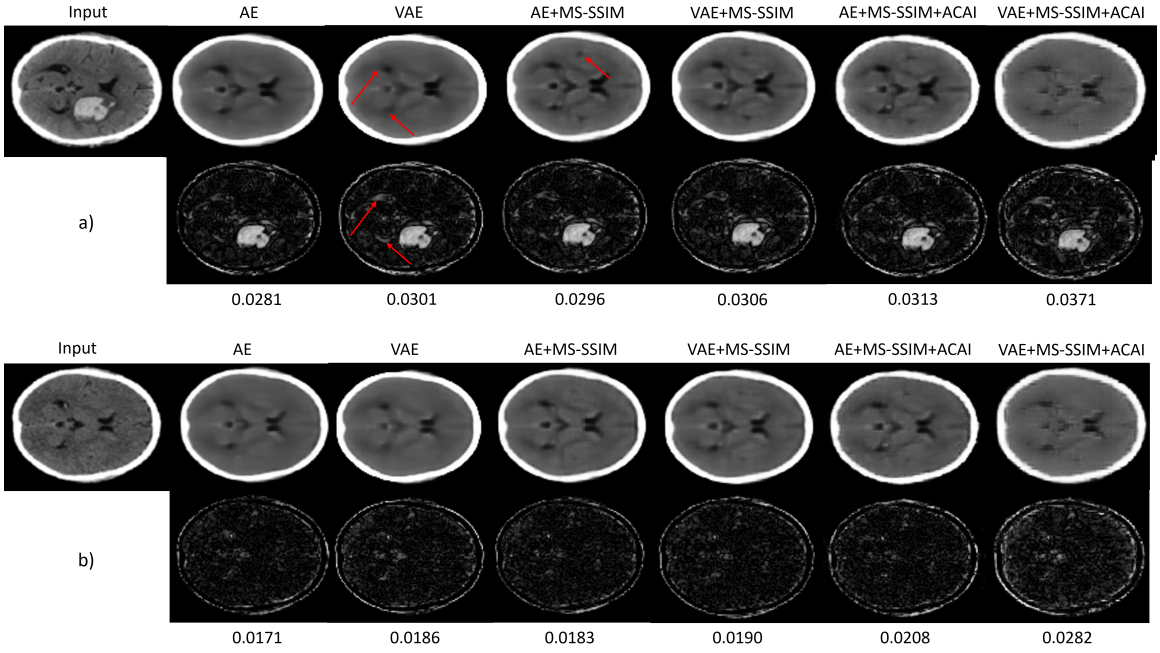

Figure 8: Comparison on the reconstruction quality for all methods compared and implemented.

Table 3: The comparison of AUROC scores between a VAE and the same VAE trained on skull-stripped data for different pathologies.

Pathology (n. scans)	**Method**	***L1***		***L1+MS-SSIM***		***L1+MS-SSIM+ACAI***	
		AUROC(I)	AUROC(F)	AUROC(I)	AUROC(F)	AUROC(I)	AUROC(F)
Atrophy	VAE	0.722	0.517	0.883	0.483	0.894	0.458
(15)	StripVAE	0.767	0.389	0.950	0.694	**0.967**	0.783
Fracture	VAE	**0.734**	0.465	0.712	0.365	0.596	0.479
(26)	StripVAE	0.401	0.497	0.564	0.397	0.641	0.346
ICH	VAE	0.670	0.441	0.768	0.427	0.731	0.509
(155)	StripVAE	0.597	0.424	0.773	0.410	**0.789**	0.383
Ischemia	VAE	0.778	0.533	0.843	0.533	0.822	0.511
(59)	StripVAE	0.759	0.411	**0.897**	0.620	0.880	0.682
Mass	VAE	0.500	0.458	0.592	0.558	0.525	0.513
(10)	StripVAE	0.542	0.517	0.625	0.492	**0.758**	0.517
Other	VAE	0.622	0.450	0.650	0.456	0.628	0.476
(30)	StripVAE	0.589	0.478	0.667	0.606	**0.700**	0.600
TOTAL	VAE	0.673	0.461	0.756	0.451	0.724	0.506
	StripVAE	0.623	0.434	0.777	0.475	**0.790**	0.469