Medical Image Analysis 69 (2021) 101952

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Survey Paper

Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study

Christoph Baur[a,*], Stefan Denner[a], Benedikt Wiestler[c], Nassir Navab[a,d], Shadi Albarqouni[a,b,1]

[a] Chair for Computer Aided Medical Procedures (CAMP), Technical University of Munich, Boltzmannstr. 3, Garching, Germany
[b] Helmholtz AI, Helmholtz Center Munich, Ingolstädter Landstraße 1, Neuherberg, Germany
[c] Neuroradiology Department of Klinikum Rechts der Isar, Ismaningerstr. 22, Munich, Germany
[d] Whiting School of Engineering, Johns Hopkins University, Baltimore, United States

ARTICLE INFO

Article history:
Received 8 April 2020
Revised 23 December 2020
Accepted 28 December 2020
Available online 2 January 2021

Keywords:
Anomaly segmentation
Detection
Unsupervised
Brain MRI
Autoencoder
Variational
Adversarial
Generative
VAE-GAN
VAEGAN

ABSTRACT

Deep unsupervised representation learning has recently led to new approaches in the field of Unsupervised Anomaly Detection (UAD) in brain MRI. The main principle behind these works is to learn a model of normal anatomy by learning to compress and recover healthy data. This allows to spot abnormal structures from erroneous recoveries of compressed, potentially anomalous samples. The concept is of great interest to the medical image analysis community as it i) relieves from the need of vast amounts of manually segmented training data—a necessity for and pitfall of current supervised Deep Learning—and ii) theoretically allows to detect arbitrary, even rare pathologies which supervised approaches might fail to find. To date, the experimental design of most works hinders a valid comparison, because i) they are evaluated against different datasets and different pathologies, ii) use different image resolutions and iii) different model architectures with varying complexity. The intent of this work is to establish comparability among recent methods by utilizing a single architecture, a single resolution and the same dataset(s). Besides providing a ranking of the methods, we also try to answer questions like i) how many healthy training subjects are needed to model normality and ii) if the reviewed approaches are also sensitive to domain shift. Further, we identify open challenges and provide suggestions for future community efforts and research directions.

1. Introduction

MR imaging of the brain is at the heart of diagnosis and treatment of neurological diseases. When sifting MR scans, Radiologists intuitively rely on a learned model of normal brain anatomy to detect pathologies. However, reading and interpreting MR scans is an intricate process: It is estimated that in 5–10% of scans, a relevant pathology is missed (Bruno et al., 2015). Recent breakthroughs in machine learning have led to automated medical image analysis methods which achieve great levels of performance in the detection of tumors or lesions arising from neuro-degenerative diseases such as Alzheimers or Multiple Sclerosis (MS). Despite all their outstanding performances, these methods—mainly based on Supervised Deep Learning—carry some disadvantages: 1) their training calls for large and diverse annotated datasets, which are scarce and costly to obtain; 2) the resulting models are limited to the discovery of lesions which are similar to those in the training data. This is especially crucial for rare diseases, for which collecting training data poses a great challenge. Lately, there have been some Deep Learning-driven attempts towards automatic brain pathology detection which tackle the problem from the perspective of so-called Unsupervised Anomaly Detection (UAD). These approaches are more similar to how Radiologists read MR scans, do not require data with pixel-level annotations and have the potential to detect arbitrary anomalies without a-priori knowing about their appearances.

UAD has a long history in medical image analysis and in brain imaging in particular. Traditional methods are based on statistical modeling, content-based retrieval, clustering or outlier-detection (Taboada-Crispi et al., 2009). Methods specifically designed for White Matter lesion detection in brain MRI relied on adaptive thresholding techniques followed by a 3D connected component analysis (Iheme et al., 2013), Markov Random Fields (Van Leemput et al., 2001), voxel-wise k-Nearest-

* Corresponding author.
E-mail address: c.baur@tum.de (C. Baur).
[1] Shared senior authors.

https://doi.org/10.1016/j.media.2020.101952
1361-8415/

Neighbor classification (Anbeek et al., 2004; Griffanti et al., 2016) or fuzzy clustering algorithms (Shiee et al., 2010). For brain tumor segmentation, approaches primarily relied on registration techniques and ATLASes: (Chen et al., 1999) uses 3D deformable registration to an ATLAS of healthy brain anatomy for detecting and segmenting tumors, (Prastawa et al., 2004) combines robust estimation techniques with registration to an ATLAS of normal anatomy, and (Menze et al., 2010) utilizes probabilistic ATLASes to build a generative model for tumor segmentation. The task of Multiple Sclerosis lesion segmentation has been tackled with Expectation-Maximization-based approaches fused with different priors (Jain et al., 2015), Bayesian Markov-Random-Fields (Schmidt, 2017) or Dictionary Learning techniques (Weiss et al., 2013). A general review on classical anomaly detection approaches with a focus on brain CT imaging is given by Taboada-Crispi et al. (2009).

Since the rise of Deep Learning, a plethora of new, data-driven approaches has appeared. Initially, Autoencoders (AEs), with their ability to learn non-linear transformations of data onto a low-dimensional manifold, have been leveraged for cluster-based anomaly detection. Lately, a variety of works used AEs and generative modeling to not simply detect, but localize and segment anomalies directly in image-space from imperfect reconstructions of input images, which is surveyed in this work in the context of brain MRI.

The underlying idea thereby is to model the distribution of healthy anatomy of the human brain with the help of deep (generative) representation learning. Once trained, anomalies can be detected as outliers from the modeled, normative distribution. AEs (Atlason et al., 2019; Baur et al., 2018) and their generative siblings (Baur et al., 2018; Chen and Konukoglu, 2018; Pawlowski et al., 2018; Zimmerer et al., 2019; 2018) have emerged as a popular framework to achieve this by essentially learning to compress and reconstruct MR data of healthy anatomy. The respective methods can essentially be divided into two categories: 1) Reconstruction-based approaches compute a pixel-wise discrepancy between input samples and their feed-forward reconstructions to determine anomalous lesions directly in image-space; 2) Restoration-based methods (Schlegl et al., 2017; You et al., 2019) try to alter an input image by moving along the latent manifold until a normal counterpart to the input sample is found, which in turn is used again to detect lesions from the pixel-wise discrepancy of the input data and its healthy restoration. To date—albeit all of these methods report promising performances—results can hardly be compared and drawing general conclusions on their strengths & weaknesses is barely possible. This is hindered by the following issues: i) most of the works rely on very different datasets with barely overlapping characteristics for their evaluation, ii) are evaluated against different pathologies, iii) operate on different resolutions and iv) utilize different model architectures with varying model complexity. The main intent of this work is to establish comparability among a broad selection of recent methods by utilizing—where applicable—a single network architecture, a single resolution and the same dataset(s).

Contribution|Here, we provide a comparative study of recent Deep-Learning based UAD approaches for brain MRI. We compare various reconstruction- as well as restoration based methods against each other on a variety of different MR datasets with different pathologies[2]. The models are tested on four different datasets for detecting two different pathologies. To evaluate the methods without having to make general assumptions about what constitutes a detection, we utilize pixel-wise segmentation measures as a tight proxy for UAD performance. For a fair comparison, we determined a single, unified architecture on which all the methods rely in this study. This ensures that model complexity is the same for all approaches, if applicable. The performances of the originally proposed networks are also presented. Further, we provide insights on the number of healthy training samples and their impact on model performance, and peek at generalization capabilities of AE models.

[2] The code is publicly available at https://github.com/StefanDenn3r/unsupervised_anomaly_detection_brain_mri.

2. Unsupervised deep representation learning for anomaly detection

2.1. Modeling healthy anatomy

The core concept behind the reviewed methods is the modeling of healthy anatomy with unsupervised deep (generative) representation learning. Therefor, the methods leverage a set of healthy MRI scans $\mathcal{X}_{healthy} \in \mathcal{R}^{D\times H\times W}$ and learn to project it to and recover it from a lower dimensional distribution $\mathbf{z} \in \mathcal{R}^K$ (see Fig. 1). In the following, we first shed the light on the ways how this normative distribution can be modeled, and then present different approaches how anomalies can be discovered using trained models.

Autoencoders|Early work in this field relied on classic AEs (Fig. 2a) to model the normative distribution: An encoder network $Enc_\theta(\mathbf{x})$ with parameters θ is trained to project a healthy input sample $\mathbf{x} \in \mathcal{X}_{healthy}$ to a lower dimensional manifold $\mathbf{z}$, from which a decoder $Dec_\phi(\mathbf{z})$ with parameters ϕ then tries to reconstruct the input as $\hat{\mathbf{x}} = Dec_\phi(Enc_\theta(\mathbf{x}))$. In other words, the model is trained to compress and reconstruct healthy anatomy by minimizing a reconstruction loss $\mathcal{L}$

$$arg\,min_{\phi,\theta}\mathcal{L}_{AE}^{\phi,\theta}(\mathbf{x},\hat{\mathbf{x}}) = \mathcal{L}_{Rec}^{\phi,\theta}(\mathbf{x},\hat{\mathbf{x}}) = \ell_1(\mathbf{x},\hat{\mathbf{x}}), \quad (1)$$

which in our case is the ℓ_1-distance between input and reconstruction. The rationale behind this is the assumption that an AE trained on only healthy samples cannot properly reconstruct anomalies in pathological data. This approach has been successfully applied to anomaly segmentation in brain MRI (Atlason et al., 2019; Baur et al., 2018) and in head CT (Sato et al., 2018). A slightly different attempt was made by Zimmerer et al. (2018), where the reconstruction-problem was turned into an inpainting-task using a Context Autoencoder (Context AE) (Fig. 2e), in which the model is trained to recover missing sections in healthy training images. The natural choice for the shape of $\mathbf{z}$, here also referred to as *latent space, bottleneck* or *manifold*, is a 1D vector. However, it has been shown that spatial AEs with a tensor-shaped bottleneck can be beneficial for high-resolution brain MRI as they preserve spatial context and can generate higher quality reconstructions (Baur et al., 2018).

Latent Variable Models|In classic AEs, there is no regularization on the manifolds structure. Constrained AEs (Chen and Konukoglu, 2018) enforce the latent representation of a training input sample to be similar to the latent representation of its reconstruction, which has proven beneficial for anomaly detection. Another way of regularizing the manifold structure can be seen in latent variable models such as Variational Autoencoders (VAEs, (Kingma and Welling, 2014), Fig. 2b), which constrain the latent space by leveraging the encoder and decoder networks of AEs to parameterize an approximation to the posterior distribution $q(\mathbf{z}) \sim \mathcal{N}(\mathbf{z}_\mu, \mathbf{z}_\sigma)$ in the latent space, using the following objective:

$$\begin{aligned} arg\,min_{\phi,\theta}\mathcal{L}_{VAE}^{\phi,\theta}(\mathbf{x},\hat{\mathbf{x}}) &= \mathcal{L}_{Rec}^{\phi,\theta}(\mathbf{x},\hat{\mathbf{x}}) + \lambda_{KL}\mathcal{L}_{KL}^{\theta}(q(\mathbf{z}),p(\mathbf{z})) \\ &= \ell_1(\mathbf{x},\hat{\mathbf{x}}) + \lambda_{KL}\mathcal{D}_{KL}(q(\mathbf{z})||p(\mathbf{z})), \end{aligned}$$

where λ_{KL} is a Lagrangian multiplier which weights the reconstruction loss against the distribution-matching KL-Divergence $\mathcal{D}_{KL}(\cdot||\cdot)$ (the original paper does not use a λ_{KL}, i.e. $\lambda_{KL} = 1$).

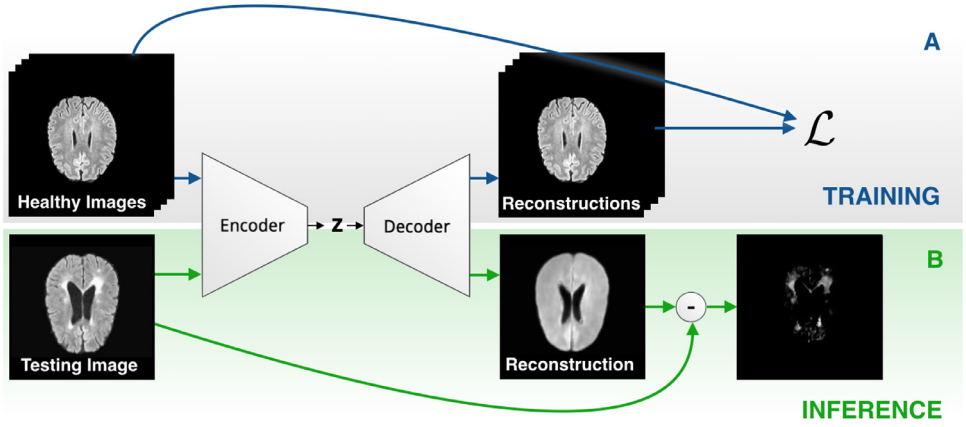


Fig. 1. The concept of Autoencoder-based Anomaly Detection/Segmentation: A) Training a model from only healthy samples and B) anomaly segmentation from erroneous reconstructions of input samples, which might carry an anomaly.

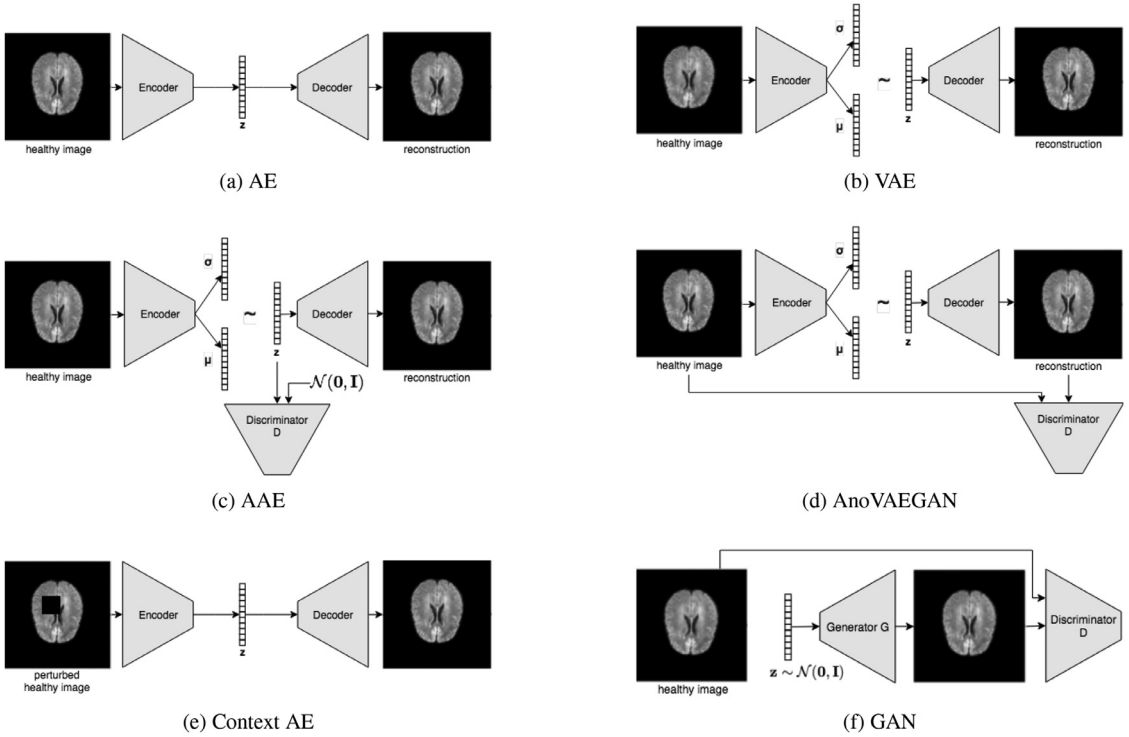
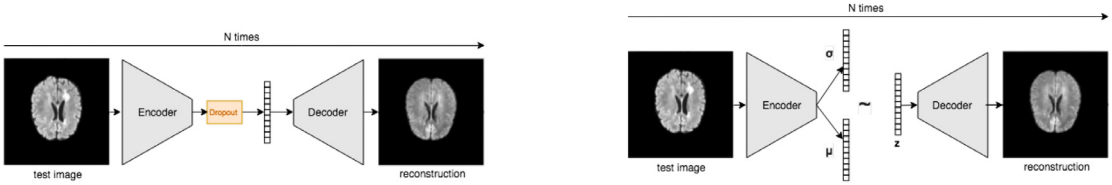


Fig. 2. Autoencoder-based architectures for UAD at a glance.

In practice, the VAE projects input data onto a learned mean μ and variance σ that parameterize a distribution, from which a sample is drawn and then reconstructed (see Fig. 2b). While the VAE tries to match $q(\mathbf{z})$ to a prior $p(\mathbf{z})$ (typically a multivariate normal distribution) by minimizing the KL-Divergence, which has various shortcomings, the so-called Adversarial Autoencoder (AAE (Makhzani et al., 2016), Fig. 2c) leverages an adversarial network as a proxy metric to minimize this discrepancy between the learned distribution $q(\mathbf{z})$ and the prior $p(\mathbf{z})$. One of the limitations of the KL-divergence is the fact that it encourages the posterior to capture the mode of the prior, but not necessarily the whole distribution (Makhzani et al., 2016; Ghosh et al., 2019). As opposed to the KL-Divergence, the optimization via an adversarial network does not favor modes of distributions and is always differentiable. Another extension to the VAE, the so-called Gaussian Mixture VAE (GMVAE, (Dilokthanakul et al., 2016)) even replaces the uni-modal prior of the VAE with a gaussian mixture, leading to higher expressive power. Due to their ability to model the underlying distribution of high dimensional data, these frameworks are naturally suited for modeling the desired normative distribution. Further, their probabilistic nature facilitates the development of principled density-based anomaly detection methods. Consequently, they have been widely employed for outlier-based anomaly detection: VAEs were used in brain MRI for MS lesion (Baur et al., 2018), tumor and stroke detection (Zimmerer et al., 2018). They have also been utilized for tumor detection in head CT (Pawlowski et al., 2018) from aggregate means of Monte-Carlo reconstructions. In brain MRI, AAE- (Chen and Konukoglu, 2018) and GMVAE-based (You et al., 2019) approaches have also been successfully employed for tumor detection.

Generative Adversarial Networks|Pioneering work, even before AEs were successfully applied for UAD in medical imaging, leveraged Generative Adversarial Networks (GANs, (Goodfellow et al., 2014), Fig. 2d) to detect anomalies in OCT data. Therefor, (Schlegl et al., 2017) modeled the distribution of healthy retinal patches with GANs and determined anomalies by computing the



(a) Bayesian AE (b) Bayesian VAE

Fig. 3. Monte Carlo Reconstructions aggregate and average N reconstructions for a single sample.

discrepancy between the retinal patch and a healthy counterpart restored by the GAN. Inspired by this work, (Baur et al., 2018) leveraged the VAEGAN (Larsen et al., 2015)—a combination of the GAN and VAE (Fig. 2d)—to overcome the training instabilities of the GAN and to allow for faster feed-forward inference, which they successfully employed for anomaly segmentation in brain MRI. In recent follow-up work, (Schlegl et al., 2019) improved on their GAN and also introduced an efficient way to replace the costly iterative restoration method by a single forward pass through the network. The concept involves the training of a Wasserstein GAN on a healthy dataset, freezing the weights of the generator and using it as the decoder of an AE; The encoder of this AE is then trained with a reconstruction objective while keeping the decoders' parameters fixed.

2.2. Anomaly segmentation

The trained models can be used for anomaly detection & segmentation in a variety of ways, which are summarized in the following. The interested reader is referred to the original papers for more detailed information.

Reconstruction Based Methods|Such approaches rely on pixel-wise residuals obtained from the difference

$$\mathbf{r} = |\mathbf{x} - \hat{\mathbf{x}}| \tag{2}$$

of input samples $\mathbf{x}$ and their reconstruction $\hat{\mathbf{x}}$ (see Fig. 1). The underlying assumption is that anomalous structures, which have never been seen during training, cannot be properly reconstructed from the distribution encoded in the latent space, such that reconstruction errors will be high for anomalous structures.

Monte Carlo Methods|For non-deterministic generative models such as VAEs, multiple reconstructions can be obtained by Monte-Carlo (MC) sampling the latent space and an average consensus residual can be computed (Pawlowski et al., 2018)

$$\mathbf{r} = \frac{1}{N}\sum_{n=1}^{N} |\mathbf{x} - \hat{\mathbf{x}}_\mathbf{n}|, \tag{3}$$

with N being the number of MC samplings and $\hat{\mathbf{x}}_\mathbf{n}$ being a single MC reconstruction. For deterministic AEs, a similar effect can be achieved by applying dropout with rate p_r to the latent space during inference time, which is also investigated in this work (see Fig. 3a and Fig. 3b for a visual explanation).

Gradient-Based Methods|The gradient-based method proposed by (Zimmerer et al., 2018) solely relies on image gradients obtained from a single backpropagation step when virtually optimizing for the following objective,

$$\begin{aligned} &arg\,min_{\tilde{\mathbf{x}}} \mathcal{L}_{Rec}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda_{KL}\mathcal{L}_{KL}(\mathbf{z}, p(\mathbf{z})) \\ &\quad = \ell_1(\mathbf{x}, \hat{\mathbf{x}}) + \lambda_{KL}\mathcal{D}_{KL}(\mathbf{z}||p(\mathbf{z})) \end{aligned} \tag{4}$$

i.e. the pursuit of bringing the reconstruction $\hat{\mathbf{x}}$ and input $\mathbf{x}$ of the model together while simultaneously moving the latent representation of an input sample closer to the prior (the normal distribution). The resulting pixel-wise gradients are used as a saliency map for anomalies, where it is assumed that stronger gradients constitute anomalies.

Restoration Based Methods|In contrast to reconstruction based methods, restoration based methods involve an optimization on the latent manifold. In the pioneering approach using GANs (Schlegl et al., 2017), the goal is to iteratively move along the GANs input distribution $\mathbf{z}$ until a healthy variant of a query image is reconstructed well. Similarly, the method from You et al. (2019) tries to restore a healthy counterpart $\hat{\mathbf{x}}$ of an input sample $\mathbf{x}$, but by altering it until the likelihood of its latent representation $\mathbf{z}$ is maximized. This can be achieved by initializing $\hat{\mathbf{x}} = \mathbf{x}$ and then iteratively optimizing $\hat{\mathbf{x}}$ for the objective in Eq. 4. Again, the anomalies can be detected in image space from residual maps $\mathbf{r}$ (see Eq. 2).

3. Experiments

In the following, we first introduce the datasets used in the experiments, together with their pre-processing, and then introduce the unified network architecture which is the foundation of all the subsequently investigated models. We further explain our post-processing pipeline and all the metrics used in our investigations, before we finally present and discuss the results from various perspectives.

3.1. Datasets

For this survey, we rely on three different datasets. Selection criteria for these datasets were i) the availability of corresponding T1, T2 and FLAIR scans per subject to be able to leverage a single shared preprocessing pipeline and ii) each dataset being produced with a different MR device.

Healthy, MS & GB|The primary dataset used in this comparative study is a homogenous set of MR scans of both healthy and diseased subjects, produced with a single Philips Achieva 3T MR scanner. It comprises FLAIR, T2- and T1-weighted MR scans of 138 healthy subjects, 48 subjects with MS lesions and 26 subjects with Glioma. All scans have been carefully reviewed and annotated by expert Neuro-Radiologists. Informed consent was waived by the local IRB.

MSLUB|The second MRI dataset (Lesjak et al., 2018) consists of co-registered T1, T2 and FLAIR scans of 30 different subjects with MS. Images have been acquired with a 3T Siemens Magnetom Trio MR system at the University Medical Center Ljubljana (UMCL). A gold standard segmentation was obtained from consensus segmentations of three expert raters.

MSSEG2015|The third MRI dataset in our experiments is the publicly available training set of the 2015 Longitudinal MS lesion segmentation challenge (Carass et al., 2017), which contains 21 scan sessions from 5 different subjects with T1, T2, PD and FLAIR images each. All data has been acquired with a 3.0 Tesla Philips MRI scanner. The exact device is not known, but the intensity distribution is different from our primary MS & GB datasets. Thus, in this study we utilize the data to test the generalization capabilities of the models and approaches.

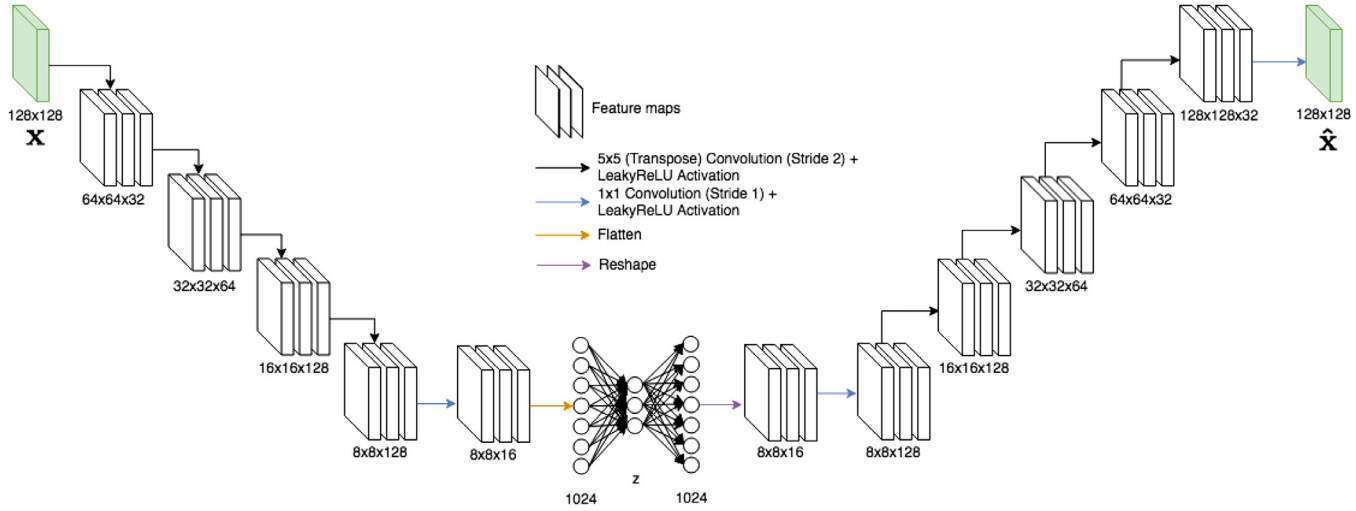


Fig. 4. The unified network architecture with a dense bottleneck. In the case of a spatial bottleneck, the flatten-, dense- and reshape-layers are replaced by a single set of 2D convolutional kernels.

Preprocessing and Split|All scans have been brought to the SRI24 ATLAS (Rohlfing et al., 2009) space to ensure all data share the same volume size and orientation. This is achieved by registering the T2 scans of all subjects to the ATLAS; The corresponding T1 and FLAIR scans are first registered to the original respective T2 volume, and then also transformed into the ATLAS space. In succession, the T1 scans in ATLAS space are utilized to determine a brainmask using ROBEX (Iglesias et al., 2011). For the experiments, we rely only on the FLAIR scans in ATLAS space after denoising them with CurvatureFlow (Sethian, 1999). Prior to feeding the data to the networks, all volumes have been normalized into the range [0,1] by dividing each scan by its 98th percentile. All datasets have randomly been split (patient-wise) into training, validation and testing sets as listed in Table 1. Training and testing is done on all axial slices of each volume for which the corresponding brainmask indicates the presence of brain pixels. Modus operandi is at a slice resolution of 128×128px. This is in stark contrast to some other works, which restrict themselves anatomically to the axial midline (Baur et al., 2018) or lower resolution (Chen and Konukoglu, 2018; Pawlowski et al., 2018).

3.2. Network architecture and models

The unified architecture depicted in Fig. 4 was empirically determined in a manual iterative architecture search by varying the number of filters and layers of an AE with a dense bottleneck. The goal of this search was to achieve a low reconstruction error on both the training and validation data from $\mathcal{D}_{healthy}$. As such, the model was not optimized for segmentation performance and thus does not favor any particular method. This unified architecture was used to train a great variety of models coming from the different, previously introduced domains.

Autoencoders|As a baseline, we used the unified architecture to train a variety of non-generative AEs:

Table 1
Training, Validation, Testing subjects of the datasets used in this study.

Dataset	Training	Validation	Testing
$\mathcal{D}_{healthy}$	110	28	-
$\mathcal{D}_{MS}$	-	3	45
$\mathcal{D}_{GB}$	-	-	28
$\mathcal{D}_{MSSEG2015}$	-	-	20
$\mathcal{D}_{MSLUB}$	-	-	30

1. **AE (dense)**: an AE with a dense bottleneck $\mathbf{z} \in \mathcal{R}^{128}$
2. **AE (spatial)** (Baur et al., 2018): an AE with a spatial bottleneck $\mathbf{z} \in \mathcal{R}^{8\times8\times128}$
3. **AE (spatial, 8 $\times$ 8$\times$2)** (Baur et al., 2018): an AE with a spatial bottleneck $\mathbf{z} \in \mathcal{R}^{8\times8\times2}$, i.e. only 128 coefficients in the bottleneck
4. **Context AE** (Zimmerer et al., 2018): with $\mathbf{z} \in \mathcal{R}^{128}$, randomly positioned 20×20 px sized masks
5. **Constrained AE** (Chen and Konukoglu, 2018): with $\mathbf{z} \in \mathcal{R}^{128}$, $\lambda = 1.0$

Latent Variable Models|Further, we trained various generative latent variable models using the same unified architecture and bottleneck configurations:

1. **VAE** (Baur et al., 2018; Zimmerer et al., 2019): with $\mathbf{z} \in \mathcal{R}^{128}$
2. **Context VAE** (Zimmerer et al., 2018): with $\mathbf{z} \in \mathcal{R}^{128}$, randomly positioned 20×20 px sized masking
3. **Constrained AAE** (Chen and Konukoglu, 2018): with $\mathbf{z} \in \mathcal{R}^{128}$, $\lambda = 1$
4. **GMVAE (dense)** (You et al., 2019): with $\mathbf{z} \in \mathcal{R}^{128}$, $c = 9$
5. **GMVAE (spatial)** (You et al., 2019): with $\mathbf{z} \in \mathcal{R}^{8\times8\times128}$, $c = 9$

Generative Adversarial Networks|Finally, we also trained an AnoVAEGAN (Baur et al., 2018) and an f-AnoGAN (Schlegl et al., 2019), whose encoder-decoder networks implement the unified architecture, and the discriminator network is a replica of the encoder:

1. **AnoVAEGAN** (Baur et al., 2018): with $\mathbf{z} \in \mathcal{R}^{128}$
2. **fAnoGAN** (Schlegl et al., 2019): with $\mathbf{z} \in \mathcal{R}^{128}$

Noteworthy, both methods were optimized with the Wasserstein loss (Arjovsky et al., 2017) to avoid GAN training instabilities and mode collapse.

All models were trained from $\mathcal{D}_{healthy}$ until convergence using an automatic early stopping criterion, i.e. training was stopped if the reconstruction loss on the held-out validation set from $\mathcal{D}_{healthy}$ did not improve more than an $\epsilon > 10e-9$ for 5 epochs. In succession, all the methods were used for reconstruction-based anomaly detection. The trained VAE and GMVAE were also used for the density-based image restoration (You et al., 2019), where each sample was restored in 500 iterations, using a learning rate of $1e-3$:

1. **VAE (restoration)** (You et al., 2019)
2. **GMVAE (restoration)** (You et al., 2019)

Table 2
Hyperparameters for the different models.

Param	Value
learning rate	0.0001
λ_{KL}	1.0
dropout rate p_r	0.2

Both AE (dense) and VAE were also used for MC-reconstruction based anomaly detection:

1. **Bayesian AE** (Pawlowski et al., 2018): Dropout rate 0.2
2. **Bayesian VAE** (Pawlowski et al., 2018): $N = 100$ MC-samples per input slice

and in the case of the Context VAE, we also tried the gradient-based approach proposed in Zimmerer et al. (2018):

1. **Context VAE (gradient)** (Zimmerer et al., 2018)

Methodology-specific hyperparameters were taken from the papers in which they were proposed. Additional hyperparameters can be taken from Table 2 and a small ablation on other lagrangian multipliers is provided in Table 9

3.3. Postprocessing

The output of all models and approaches is subject to the same post-processing. Every residual image $\mathbf{r}$ is first multiplied with a slightly eroded brain-mask to remove prominent residuals occurring near sharp edges at brain-mask boundaries and gyri and sulci (the latter are very diverse and hard to model). Further, for the MS lesion datasets we make use of prior knowledge and only keep positive residuals as these lesions are known to be fully hyperintense in FLAIR images. For each MR volume, the residual images for all slices are first aggregated into a corresponding 3D residual volume, which is then subject to a 3D median filtering with a $5 \times 5 \times 5$ kernel to remove small outliers and to obtain a more continuous signal. The latter is beneficial for the subsequent model assessment as it leads to smoother curves (see Subsection 3.18 for an analysis of the impact of post-processing on model performance). As a final step, the continuous output is binarized and a 3D connected component analysis is performed on the resulting binary volumes to discard any small structures with an area less than 8 voxels.

3.4. Metrics

We assess the anomaly segmentation performance at a level of single voxels, at which class imbalance needs careful consideration as anomalous voxels are usually less frequent than normal voxels. To do so, we generate dataset-specific Precision-Recall-Curves (PRC) and then compute the area under it (AUPRC). Noteworthy, this allows to judge the models capabilities without choosing an Operating Point (OP). Further, for each model we provide an estimate of its theoretically best possible DICE-score ($\lceil \text{DICE} \rceil$) on each dataset. Therefor, for each testing dataset $d \in \mathcal{D} = \mathcal{D}_{MS}, \mathcal{D}_{GB}, \mathcal{D}_{MSSEG2015}, \mathcal{D}_{MSLUB}$, we utilize the available ground-truth segmentation and perform a greedy search up to three decimals to determine the respective OP on the PRC curve which yields the best possible DICE score for dataset d. Additionally, to simulate the models performance in more realistic settings, we utilize a held-out validation set from $\mathcal{D}_{MS}$ to determine an OP t at which we then compute patient-specific DICE-scores for every dataset. Some of the reviewed works originally utilize Receiver-Operating-Characteristics (ROC) to evaluate anomaly detection performance. We report the area under such ROC curves (AUROC) as well, but want to emphasize that it has to be used with care. Under heavy class imbalance, ROC curves can be misleading as they give much higher weight to the more frequent class and thus in the case of a pixel-wise assessment very optimistic views on performance. Therefore, we promote the AUPRC as a much more sensible metric.

To gain deeper insights what makes a model capable of segmenting anomalies better than others, we also report dataset-specific ℓ_1-reconstruction errors on normal (ℓ_1-RE$_N$) and anomalous voxels (ℓ_1-RE$_A$), as well as the $\mathcal{X}^2$-distance of the respective normal and anomalous residual histograms for every model. The reconstruction errors are determined for every voxel of interest, aggregated via summation and normalized by the total number of aggregated voxels.

3.5. Overview

Detailed results of all models and UAD approaches on all datasets can be found in Tables 3 ($\mathcal{D}_{MS}$), 4 ($\mathcal{D}_{GB}$), 5 ($\mathcal{D}_{MSLUB}$) and 6 ($\mathcal{D}_{MSSEG2015}$). In the following, we analyze all these data from different perspectives. We start by first comparing different model types and bottleneck design, followed by the different ways to detect anomalies directly in image-space. Then, we shed the light on the number of training subjects and their impact on performance, and elaborate on domain shift.

3.6. Constraining & regularization

Initially, we compare the classic AE (dense) to its VAE and Constrained AE counterpart to investigate the effect of constraining or regularizing the latent space of the models. Recall that VAEs regularize the latent space to follow a prior distribution, whereas the deterministic Constrained AE enforces that reconstructions and input lie closely on the manifold. We measure the models' performances in terms of the AUPRC as well as the $\lceil \text{DICE} \rceil$ and glimpse at the reconstruction errors for normal and anomalous pixels (see Fig. 11 for residual histograms of normal and anomalous voxels). We see from Table 3 that explicitly modeling a distribution with a VAE leads to dramatic performance gains on $\mathcal{D}_{MS}$ over the standard AE, and introducing the matching constraint (Constrained AE) between $\mathbf{x}$ and $\hat{\mathbf{x}}$ improves the performance even more. On all other datasets, the VAE clearly is the winner among the compared models, but the Constrained AE still outperforms the classic AE. From these results, we deduce that enforcing a structure on the manifold of AEs is indeed beneficial for UAD.

3.7. Dense vs spatial bottleneck

To determine if the design of the AE bottleneck can improve the performance of the models, we further compare dense models for which a spatial counterpart exists, i.e. AE (dense) vs AE (spatial) vs AE (spatial $8 \times 8 \times 2$) vs GMVAE (dense) vs GMVAE (spatial). The spatial bottleneck allows the model to preserve spatial information and geometric features in its latent space, which positively affects the models reconstruction capabilities. Experimenting with different numbers of channels in the spatial bottleneck, we find that at our resolution of 128x128px, the spatial models with 128 channels reconstruct their input too well (see Fig. 7), including the anomalies. This impacts anomaly segmentation performance, as Tables 3–6 show: the dense models outperform the spatial variants with more than two channels in the bottleneck on all MS datasets. Alone on $\mathcal{D}_{GB}$, AE (spatial) with its 128 channels performs slightly better than its dense counterpart. When the spatial bottleneck is designed to have the same number of coefficients (AE (spatial $8 \times 8 \times 2$)) in the bottleneck as AE (dense), tumor segmentation is improved while the performance is comparable to AE (dense) on MS data.

Table 3
Experimental results on the MS dataset.

Approach	AUPRC	AUROC	⌈DICE⌉	DICE ($\mu \pm \sigma$)	ℓ_1-RE$_N$ ($\mu \pm \sigma$)	ℓ_1-RE$_A$ ($\mu \pm \sigma$)	$\mathcal{X}^2$
AE (dense)	0.271	0.918	0.389	0.325 ± 0.164	5.30e-10 ± 5.45e-06	4.07e-08 ± 4.29e-05	3.59e-01
AE (spatial) (Baur et al. (2018))	0.13	0.852	0.231	0.165 ± 0.134	3.86e-10 ± 1.91e-06	1.78e-08 ± 1.25e-05	8.39e-02
AE (spatial 8x8x2) (Baur et al. (2018))	0.273	0.986	0.392	0.311 ± 0.178	1.60e-08 ± 4.74e-05	1.67e-07 ± 6.20e-05	4.78e-01
VAE (Baur et al. (2018); Zimmerer et al. (2019))	0.399	0.945	0.469	0.389 ± 0.166	8.27e-10 ± 8.11e-06	7.30e-08 ± 6.32e-05	4.46e-01
VAE (restoration) (You et al. (2019))	0.454	0.946	0.495	0.404 ± 0.176	8.92e-10 ± 8.49e-06	7.77e-08 ± 6.63e-05	4.61e-01
Context AE (Zimmerer et al. (2018))	0.233	0.9	0.374	0.327 ± 0.173	6.27e-10 ± 6.78e-06	6.68e-08 ± 6.01e-05	4.26e-01
Context VAE (Zimmerer et al. (2018))	0.416	0.937	0.492	0.418 ± 0.167	6.09e-10 ± 6.27e-06	7.15e-08 ± 5.93e-05	4.17e-01
Context VAE (gradient) (Zimmerer et al. (2018))	0.294	0.963	0.385	0.305 ± 0.161	9.00e-10 ± 8.38e-06	3.18e-08 ± 4.68e-05	2.77e-01
GMVAE (dense) (You et al. (2019))	0.389	0.944	0.477	0.387 ± 0.178	8.11e-10 ± 8.08e-06	7.60e-08 ± 6.44e-05	4.51e-01
GMVAE (dense restoration) (You et al. (2019))	0.453	0.945	0.501	0.411 ± 0.180	9.07e-10 ± 8.64e-06	8.34e-08 ± 6.84e-05	4.77e-01
GMVAE (spatial) (You et al. (2019))	0.096	0.877	0.191	0.148 ± 0.126	3.98e-10 ± 1.91e-06	1.57e-08 ± 1.02e-05	7.91e-02
GMVAE (spatial restoration) (You et al. (2019))	0.295	0.925	0.363	0.287 ± 0.157	2.81e-10 ± 2.04e-06	1.60e-08 ± 1.31e-05	1.12e-01
f-AnoGAN (Schlegl et al. (2019))	0.448	0.957	0.489	0.417 ± 0.178	2.05e-09 ± 1.54e-05	1.06e-07 ± 7.28e-05	4.96e-01
AnoVAEGAN (Baur et al. (2018))	0.376	0.947	0.45	0.371 ± 0.178	1.22e-09 ± 1.10e-05	9.78e-08 ± 7.37e-05	5.03e-01
Constrained AE (Chen and Konukoglu (2018))	0.429	0.94	0.485	0.409 ± 0.173	5.73e-10 ± 5.68e-06	4.16e-08 ± 4.29e-05	3.64e-01
Constrained AAE (Chen and Konukoglu (2018))	0.268	0.949	0.392	0.331 ± 0.195	1.66e-09 ± 1.42e-05	1.04e-07 ± 7.94e-05	5.06e-01
Bayesian AE (Pawlowski et al. (2018))	0.262	0.913	0.373	0.313 ± 0.159	5.44e-10 ± 5.56e-06	4.23e-08 ± 4.36e-05	3.72e-01
Bayesian VAE (Pawlowski et al. (2018))	0.403	0.945	0.471	0.390 ± 0.165	8.23e-10 ± 8.09e-06	7.31e-08 ± 6.37e-05	4.46e-01

Table 4
Experimental results on the GB dataset.

Approach	AUPRC	AUROC	⌈DICE⌉	DICE ($\mu \pm \sigma$)	ℓ_1-RE$_N$ ($\mu \pm \sigma$)	ℓ_1-RE$_A$ ($\mu \pm \sigma$)	$\mathcal{X}^2$
AE (dense)	0.158	0.753	0.299	0.268 ± 0.133	1.73e-09 ± 1.14e-05	6.05e-08 ± 6.57e-05	4.20e-01
AE (spatial) (Baur et al. (2018))	0.179	0.737	0.295	0.239 ± 0.127	7.82e-10 ± 2.78e-06	2.88e-08 ± 1.80e-05	4.08e-02
AE (spatial 8x8x2) (Baur et al. (2018))	0.266	0.936	0.358	0.300 ± 0.174	1.69e-08 ± 4.68e-05	1.49e-07 ± 8.78e-05	4.82e-01
VAE (Baur et al. (2018); Zimmerer et al. (2019))	0.272	0.795	0.441	0.374 ± 0.162	3.62e-09 ± 2.17e-05	1.52e-07 ± 1.34e-04	6.18e-01
VAE (restoration) (You et al. (2019))	0.441	0.8	0.537	0.435 ± 0.193	2.96e-09 ± 1.73e-05	1.97e-07 ± 1.55e-04	6.45e-01
Context AE (Zimmerer et al. (2018))	0.253	0.753	0.402	0.343 ± 0.160	1.69e-09 ± 1.18e-05	1.23e-07 ± 9.80e-05	4.28e-01
Context VAE (Zimmerer et al. (2018))	0.215	0.775	0.375	0.333 ± 0.139	2.27e-09 ± 1.48e-05	1.03e-07 ± 9.20e-05	5.09e-01
Context VAE (gradient) (Zimmerer et al. (2018))	0.172	0.799	0.315	0.281 ± 0.122	2.85e-09 ± 1.76e-05	5.03e-08 ± 8.02e-05	3.96e-01
GMVAE (dense) (You et al. (2019))	0.367	0.798	0.492	0.406 ± 0.176	2.98e-09 ± 1.79e-05	1.62e-07 ± 1.38e-04	6.21e-01
GMVAE (dense restoration) (You et al. (2019))	0.423	0.797	0.522	0.421 ± 0.190	2.95e-09 ± 1.74e-05	2.04e-07 ± 1.57e-04	6.48e-01
GMVAE (spatial) (You et al. (2019))	0.119	0.737	0.258	0.216 ± 0.125	8.39e-10 ± 3.00e-06	2.19e-08 ± 1.43e-05	5.39e-02
GMVAE (spatial restoration) (You et al. (2019))	0.21	0.752	0.313	0.272 ± 0.128	6.73e-10 ± 3.13e-06	2.06e-08 ± 1.89e-05	9.84e-02
f-AnoGAN (Schlegl et al. (2019))	0.349	0.786	0.447	0.379 ± 0.174	5.20e-09 ± 2.54e-05	2.32e-07 ± 1.78e-04	6.92e-01
AnoVAEGAN (Baur et al. (2018))	0.334	0.774	0.485	0.385 ± 0.191	3.65e-09 ± 2.21e-05	2.33e-07 ± 1.75e-04	6.77e-01
Constrained AE (Chen and Konukoglu (2018))	0.23	0.772	0.353	0.318 ± 0.145	1.80e-09 ± 1.14e-05	7.02e-08 ± 7.35e-05	4.69e-01
Constrained AAE (Chen and Konukoglu (2018))	0.365	0.793	0.481	0.392 ± 0.183	4.12e-09 ± 2.24e-05	2.33e-07 ± 1.75e-04	6.86e-01
Bayesian AE (Pawlowski et al. (2018))	0.143	0.747	0.28	0.253 ± 0.124	1.77e-09 ± 1.16e-05	5.81e-08 ± 6.42e-05	4.15e-01
Bayesian VAE (Pawlowski et al. (2018))	0.271	0.795	0.44	0.374 ± 0.162	3.70e-09 ± 2.21e-05	1.53e-07 ± 1.35e-04	6.18e-01

Table 5
Experimental results on the MSLUB dataset.

Approach	AUPRC	AUROC	⌈DICE⌉	DICE ($\mu \pm \sigma$)	ℓ_1-RE$_N$ ($\mu \pm \sigma$)	ℓ_1-RE$_A$ ($\mu \pm \sigma$)	$\mathcal{X}^2$
AE (dense)	0.163	0.794	0.271	0.181 ± 0.168	9.21e-10 ± 7.01e-06	4.58e-08 ± 5.42e-05	4.35e-01
AE (spatial) (Baur et al. (2018))	0.065	0.732	0.154	0.098 ± 0.116	7.05e-10 ± 2.58e-06	2.21e-08 ± 1.51e-05	1.04e-01
AE (spatial 8x8x2) (Baur et al. (2018))	0.123	0.967	0.227	0.147 ± 0.147	2.91e-08 ± 6.61e-05	2.80e-07 ± 9.38e-05	6.13e-01
VAE (Baur et al. (2018) Zimmerer et al. (2019))	0.234	0.827	0.323	0.205 ± 0.207	1.67e-09 ± 1.15e-05	8.36e-08 ± 8.84e-05	5.53e-01
VAE (restoration) (You et al. (2019))	0.275	0.839	0.333	0.203 ± 0.209	1.92e-09 ± 1.27e-05	9.31e-08 ± 9.53e-05	5.65e-01
Context AE (Zimmerer et al. (2018))	0.19	0.771	0.28	0.193 ± 0.186	9.65e-10 ± 7.32e-06	6.10e-08 ± 6.87e-05	4.90e-01
Context VAE (Zimmerer et al. (2018))	0.226	0.805	0.316	0.204 ± 0.202	1.12e-09 ± 8.34e-06	6.75e-08 ± 7.27e-05	5.12e-01
Context VAE (gradient) (Zimmerer et al. (2018))	0.154	0.889	0.265	0.175 ± 0.173	1.56e-09 ± 1.13e-05	4.21e-08 ± 6.06e-05	3.71e-01
GMVAE (dense) (You et al. (2019))	0.234	0.832	0.316	0.202 ± 0.210	1.80e-09 ± 1.22e-05	8.65e-08 ± 8.94e-05	5.56e-01
GMVAE (dense restoration) (You et al. (2019))	0.271	0.836	0.332	0.204 ± 0.208	2.01e-09 ± 1.32e-05	9.87e-08 ± 9.74e-05	5.75e-01
GMVAE (spatial) (You et al. (2019))	0.054	0.756	0.136	0.102 ± 0.106	7.07e-10 ± 2.62e-06	1.95e-08 ± 1.31e-05	1.03e-01
GMVAE (spatial restoration) (You et al. (2019))	0.147	0.804	0.23	0.158 ± 0.149	4.89e-10 ± 2.68e-06	1.81e-08 ± 1.63e-05	1.28e-01
f-AnoGAN (Schlegl et al. (2019))	0.221	0.856	0.283	0.189 ± 0.192	4.56e-09 ± 2.39e-05	1.51e-07 ± 1.21e-04	6.26e-01
AnoVAEGAN (Baur et al. (2018))	0.193	0.823	0.282	0.180 ± 0.167	2.04e-09 ± 1.36e-05	1.01e-07 ± 9.56e-05	5.89e-01
Constrained AE (Chen and Konukoglu (2018))	0.209	0.821	0.298	0.197 ± 0.187	1.11e-09 ± 8.07e-06	4.73e-08 ± 5.61e-05	4.73e-01
Constrained AAE (Chen and Konukoglu (2018))	0.203	0.852	0.289	0.194 ± 0.207	3.35e-09 ± 1.97e-05	1.26e-07 ± 1.12e-04	5.97e-01
Bayesian AE (Pawlowski et al. (2018))	0.155	0.79	0.267	0.183 ± 0.162	9.39e-10 ± 7.17e-06	4.80e-08 ± 5.57e-05	4.43e-01
Bayesian VAE (Pawlowski et al. (2018))	0.234	0.827	0.322	0.201 ± 0.206	1.68e-09 ± 1.16e-05	8.36e-08 ± 8.84e-05	5.53e-01

3.8. Latent variable models

Next, we focus only on different latent variable model types, i.e. the VAE, GMVAE (dense) and Constrained AAE. On the MS datasets $\mathcal{D}_{MS}$, $\mathcal{D}_{MSSEG2015}$ and $\mathcal{D}_{MSLUB}$, the VAE constitutes the best among the compared models. The Constrained AAE yields lower performance than the other models—also lower than its non-generative sibling, the Constrained AE. However, on the Glioblastoma dataset, it is on par with the GMVAE, and both models significantly outperform the VAE in the detection of brain tumors. Generally, the performance of the GMVAE seems to heavily depend on the dataset rather than the pathology: On

Table 6
Experimental results on the MSSEG2015 dataset.

Approach	AUPRC	AUROC	⌈DICE⌉	DICE ($\mu \pm \sigma$)	ℓ_1-RE$_N$ ($\mu \pm \sigma$)	ℓ_1-RE$_A$ ($\mu \pm \sigma$)	χ^2
AE (dense)	0.08	0.879	0.185	0.150 ± 0.075	1.87e-09 ± 1.10e-05	6.40e-08 ± 5.41e-05	4.90e-01
AE (spatial) (Baur et al. (2018))	0.037	0.781	0.106	0.066 ± 0.073	1.11e-09 ± 3.34e-06	2.68e-08 ± 1.52e-05	1.40e-01
AE (spatial 8x8x2) (Baur et al. (2018))	0.079	0.975	0.178	0.139 ± 0.116	1.44e-07 ± 1.53e-04	1.21e-06 ± 1.29e-04	1.02e+00
VAE (Baur et al. (2018) Zimmerer et al. (2019))	0.139	0.899	0.257	0.200 ± 0.124	2.89e-09 ± 1.52e-05	1.10e-07 ± 7.43e-05	6.07e-01
VAE (restoration) (You et al. (2019))	0.202	0.905	0.272	0.211 ± 0.122	3.44e-09 ± 1.69e-05	1.18e-07 ± 7.91e-05	6.10e-01
Context AE (Zimmerer et al. (2018))	0.102	0.877	0.225	0.188 ± 0.116	1.93e-09 ± 1.16e-05	8.92e-08 ± 6.34e-05	5.49e-01
Context VAE (Zimmerer et al. (2018))	0.216	0.896	0.336	0.267 ± 0.112	1.74e-09 ± 1.00e-05	9.24e-08 ± 6.47e-05	5.28e-01
Context VAE (gradient) (Zimmerer et al. (2018))	0.081	0.923	0.173	0.127 ± 0.088	2.39e-09 ± 1.48e-05	6.07e-08 ± 5.85e-05	3.87e-01
GMVAE (dense) (You et al. (2019))	0.095	0.9	0.21	0.174 ± 0.121	2.99e-09 ± 1.64e-05	1.06e-07 ± 7.13e-05	6.05e-01
GMVAE (dense restoration) (You et al. (2019))	0.199	0.909	0.28	0.223 ± 0.124	3.98e-09 ± 1.86e-05	1.31e-07 ± 8.01e-05	6.35e-01
GMVAE (spatial) (You et al. (2019))	0.042	0.846	0.106	0.069 ± 0.073	1.10e-09 ± 3.34e-06	2.73e-08 ± 1.36e-05	1.29e-0
GMVAE (spatial restoration) (You et al. (2019))	0.097	0.873	0.178	0.118 ± 0.110	8.13e-10 ± 3.52e-06	2.51e-08 ± 1.63e-05	1.57e-01
f-AnoGAN (Schlegl et al. (2019))	0.255	0.923	0.342	0.278 ± 0.140	3.43e-08 ± 7.40e-05	7.85e-07 ± 1.98e-04	9.48e-01
AnoVAEGAN (Baur et al. (2018))	0.135	0.911	0.235	0.200 ± 0.133	5.97e-09 ± 2.55e-05	1.91e-07 ± 1.03e-04	6.93e-01
Constrained AE (Chen and Konukoglu (2018))	0.137	0.9	0.261	0.209 ± 0.100	2.26e-09 ± 1.22e-05	6.76e-08 ± 5.33e-05	5.16e-01
Constrained AAE (Chen and Konukoglu (2018))	0.092	0.917	0.204	0.190 ± 0.170	1.14e-08 ± 3.98e-05	2.94e-07 ± 1.24e-04	7.50e-01
Bayesian AE (Pawlowski et al. (2018))	0.075	0.877	0.176	0.142 ± 0.072	1.89e-09 ± 1.13e-05	6.71e-08 ± 5.46e-05	4.98e-01
Bayesian VAE (Pawlowski et al. (2018))	0.137	0.898	0.252	0.194 ± 0.117	2.87e-09 ± 1.51e-05	1.08e-07 ± 7.42e-05	6.07e-01

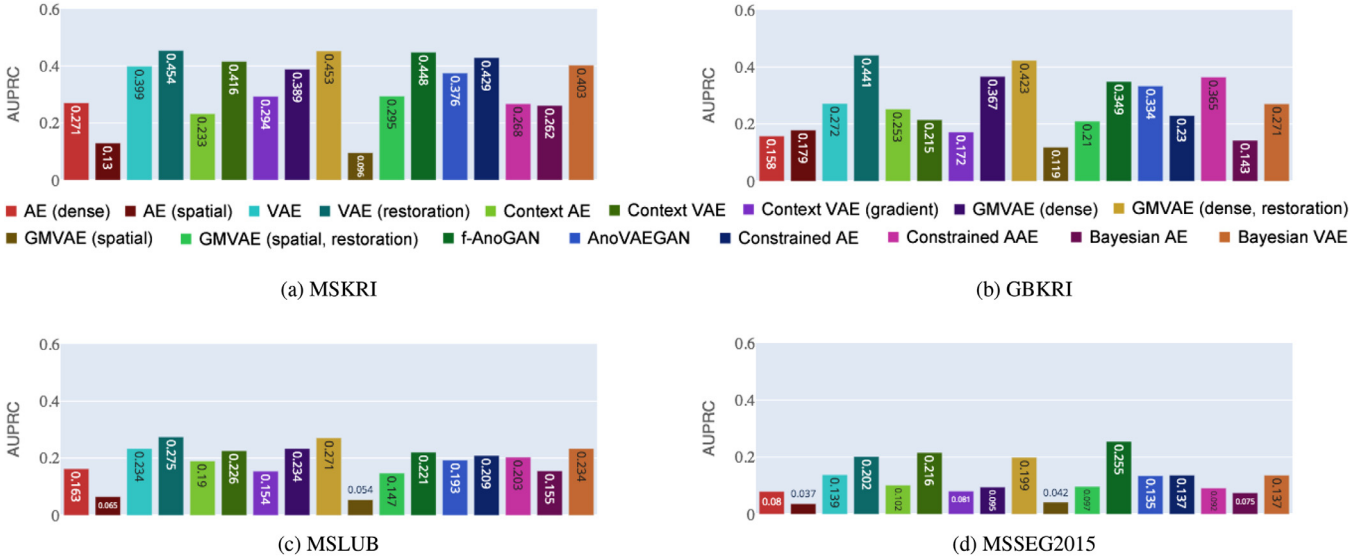
$\mathcal{D}_{MS}$ and $\mathcal{D}_{MSLUB}$ it behaves very similar to the VAE, whereas on $\mathcal{D}_{GB}$ and $\mathcal{D}_{MSSEG2015}$ its performance resembles that of the Constrained AAE.

3.9. GAN-Based models

GAN-based models are known to produce very realistic and crisp images, while AEs are known for their blurry reconstructions. Indeed, qualitative comparison of the f-AnoGAN and the AnoVAEGAN to the AE and VAE shows that the GAN-based models promote sharpness. This is particularly evident near the boundaries of the brain (see Fig. 6). However, both the f-AnoGAN and AnoVAEGAN model tend to memorize the training data, i.e. show signs of overfitting, such that reconstructions often differ anatomically from the actual input samples (see Fig. 6b for an axial midline slice from $\mathcal{D}_{MS}$). This is especially the case for the AnoVAEGAN, which produces the most crisp reconstructions, but often does not preserve anatomical coherence at all. As a result, on the MS datasets its performance is only comparable to the VAE, but it works considerably better for Glioblastoma segmentation. The f-AnoGAN does not provide as crisp images, but preserves the shape of the input sample and the difference between reconstruction residuals on normal and anomalous pixels is considerably higher across all datasets than for any of the other methods. This makes the UAD performance of the f-AnoGAN stand out. In total, both GAN-based approaches significantly outperform the standard AE (on average, more than 9% for the AnoVAEGAN and more than 15% for the f-AnoGAN) and the f-AnoGAN clearly also outperforms the VAE (on average more than 6%).

3.10. Monte-Carlo methods

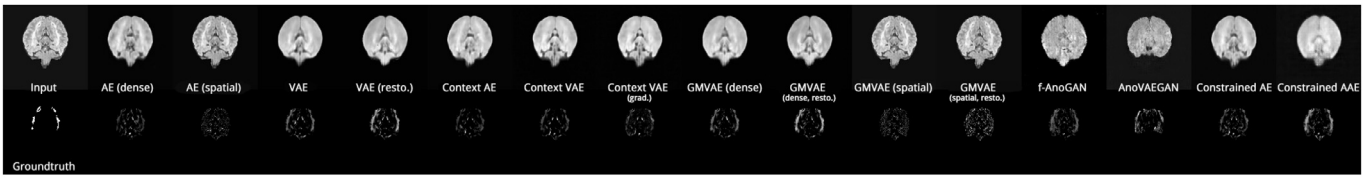
Monte-Carlo methods applied to (variational) AEs provide an interesting means to aggregate a consensus reconstruction, in which only very likely image features should be emphasized. To investigate if anomalies are affected, we experiment with $N = 100$ MC-reconstructions and—where necessary—an empirically chosen dropout-rate $p_d = 0.2$ to trade-off reconstruction quality and chance. We find that, compared to one-shot reconstructions, the impact of MC-sampling is at most subtle, and not consistent across different models and datasets. A comparison of AE (dense) to the Bayesian AE shows that MC-dropout leads to a slightly worse performance in almost all metrics across all datasets. On the other hand, the Bayesian VAE, which does not need dropout for MC



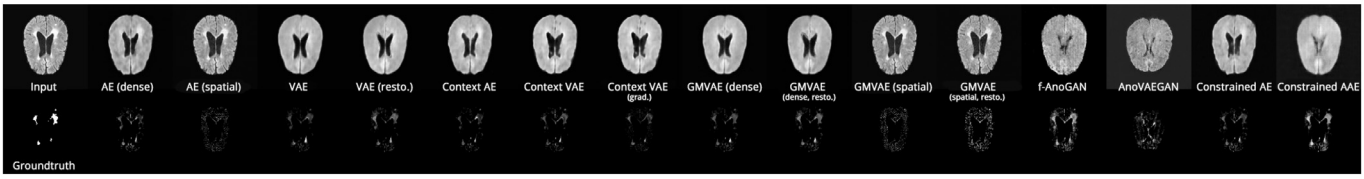
(a) MSKRI (b) GBKRI

(c) MSLUB (d) MSSEG2015

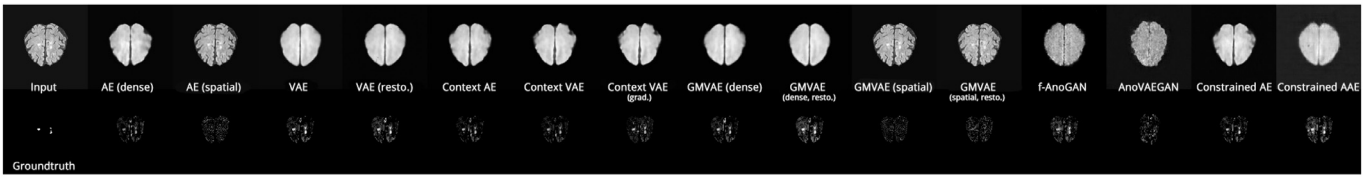
Fig. 5. AUPRC of all models and UAD approaches, using the unified architecture.



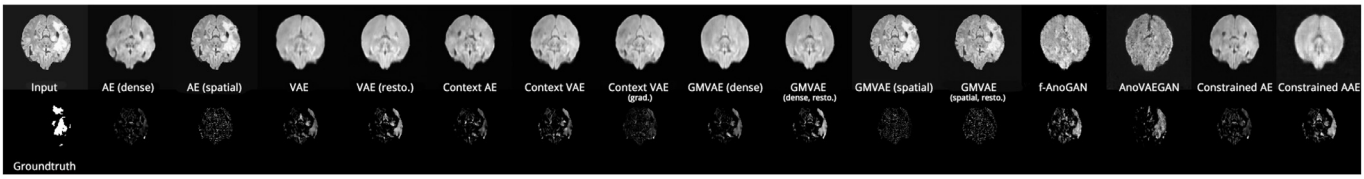
(a) axial slice from $\mathcal{D}_{MS}$, ventral



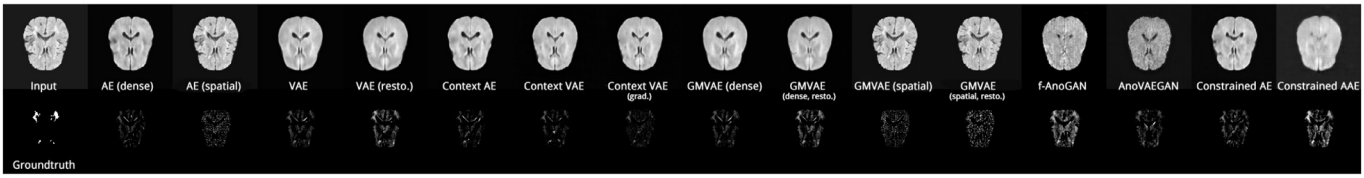
(b) axial slice from $\mathcal{D}_{MS}$, midline



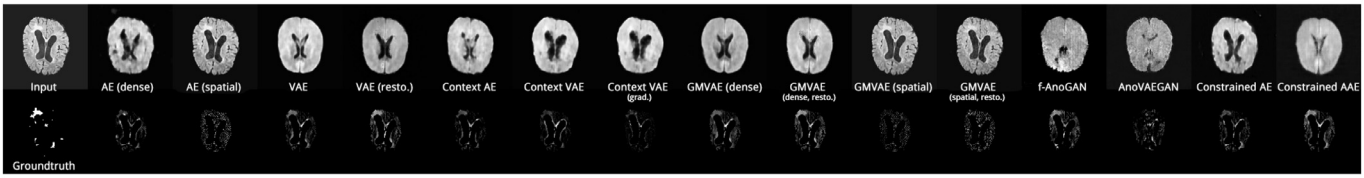
(c) axial slice from $\mathcal{D}_{MS}$, dorsal



(d) axial slice from $\mathcal{D}_{GB}$, ventral



(e) axial slice from $\mathcal{D}_{MSSEG2015}$, dorsal



(f) axial slice from $\mathcal{D}_{MSLUB}$, midline

Fig. 6. Visual examples of the different reviewed methods on different datasets, using the unified architecture. Top row: reconstructions; Bottom row: raw residuals.

sampling due to its probabilistic bottleneck, is equal to or slightly outperforms the VAE on $\mathcal{D}_{MS}$, but not on $\mathcal{D}_{GB}$ and $\mathcal{D}_{MSLUB}$. Overall, these numbers indicate that MC methods, albeit an interesting approach, do not provide significant gains in the way they are currently employed.

3.11. Reconstruction vs restoration

Previous comparisons focused on different model types and all relied on the reconstruction-based UAD concept. In the following, we rank reconstruction-based methods, against gradient- and restoration-based UAD approaches. More precisely, we compare reconstruction against restoration on the VAE, GMVAE (dense) and GMVAE (spatial). We further rank the restoration-based methods against the top-candidate f-AnoGAN. From Tables 3 to 6 it is evident that restoration based UAD is generally superior to the reconstruction-based counterparts (ranging from 4 to 17% for the VAE, 4–10% for the dense GMVAE and 6–20% for the spatial GMVAE). Consistent with our previously measured results on dense versus spatial models, we also witness a dramatic drop in performance when using the spatial GMVAE, though. Except for $\mathcal{D}_{MSSEG2015}$, the dense restoration methods outperform the f-AnoGAN in all scenarios in terms of the AUPRC and $\lceil \text{DICE} \rceil$. Overall, we consider the restoration-based methods to be preferable over the reconstruction-based approaches when run-time is not of concern.

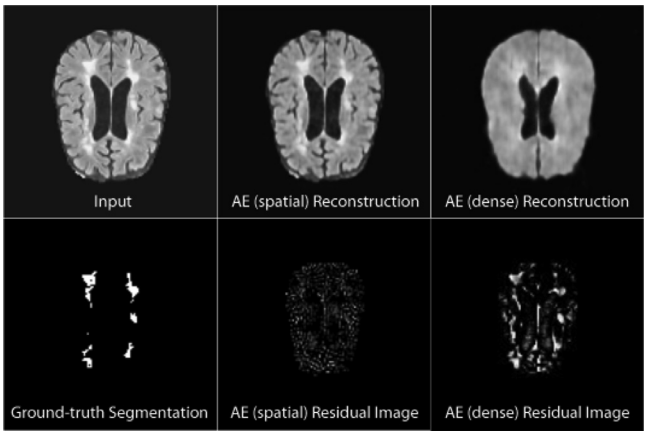


Fig. 7. Reconstructions and postprocessed residuals using dense and spatial AEs.

3.12. Domain shift

Deep Learning models trained from data coming from one domain generally have difficulties to generalize well to other domains, and tackling such domain shift is still a highly active research area. Here, we want to determine to which extent AEs are prone to this effect and if some methods generalize better than others. Subject to our investigations are the MS datasets $\mathcal{D}_{MS}$, $\mathcal{D}_{MSLUB}$ and $\mathcal{D}_{MSSEG2015}$ among which such shifts occur. Generally, UAD performance is best on $\mathcal{D}_{MS}$, which matches the training data distribution, and on both $\mathcal{D}_{MSLUB}$ & $\mathcal{D}_{MSSEG2015}$, the UAD performance drops significantly. However, the reasons for this drop can be manifold, and we want to emphasize that UAD performance as such is not a good indicator for domain shift, as the lesion size and count differs across datasets, and the contrast for $\mathcal{D}_{MS}$ is considerably better than for the other datasets. Instead, we suggest to look at the reconstruction error of normal pixels ℓ_1-RE$_N$ in these datasets. From Tables 3, 5 and 6 it can be seen that this error hardly degrades across all these datasets. This implies that generalization measured in terms of the models reconstruction capabilities is not of primary concern. However, from aforementioned tables it can be seen that the reconstruction error of anomalous pixels ℓ_1-RE$_N$ is significantly smaller on $\mathcal{D}_{MSLUB}$ and $\mathcal{D}_{MSSEG2015}$, which is a clear indicator of weaker contrast between normal tissue and lesions in these datasets. This indicates that for the reviewed approaches to work well, good contrast is required.

3.13. Different pathologies

On both Multiple Sclerosis ($\mathcal{D}_{MS}$) and Glioblastoma ($\mathcal{D}_{GB}$), the restoration-based approaches with dense bottleneck constitute the top-performers, delivering results in roughly the same league. Similarly, lowest performances can be seen from the spatial models, the gradient-based UAD approach and the standard AE. However, in contrast to $\mathcal{D}_{MS}$, on $\mathcal{D}_{GB}$ there is a large performance gap between the top-performing restoration approaches and any other methods: the GAN-based methods f-AnoGAN and AnoVAE-GAN drop by 10% and 4%, respectively, the performance of the VAE models degrades by at least 12% and the Constrained AE even loses 20% in AUPRC. Interestingly, the Constrained AAE gains by 10%. Multiple factors lead to the lower performance: In contrast to MS lesions, tumors do not purely appear hyper-intense in FLAIR MRI. Some compartments of the tumor also resemble normal tissue, and the investigated UAD approaches have difficulties to properly delineate those. Second, tumors often are not only larger than MS lesions, but can have very complex shape (see Fig. 6d). This is hard to segment with precision—even among human annotators, there is variation.

3.14. How much healthy training data is enough?

In our previous experiment, we relied on 110 healthy training subjects. The question arises whether this is a sufficient amount, or if fewer scans even lead to comparable results. To give insights into the behavior of the examined models in this context, we provide a comparison of the AUPRC of conceptually most different models, all trained at varying number of healthy subjects, i.e. 10, 50 and 100% of the available training samples. Results on the four different datasets can be seen in Fig. 8. The GAN-based mod-

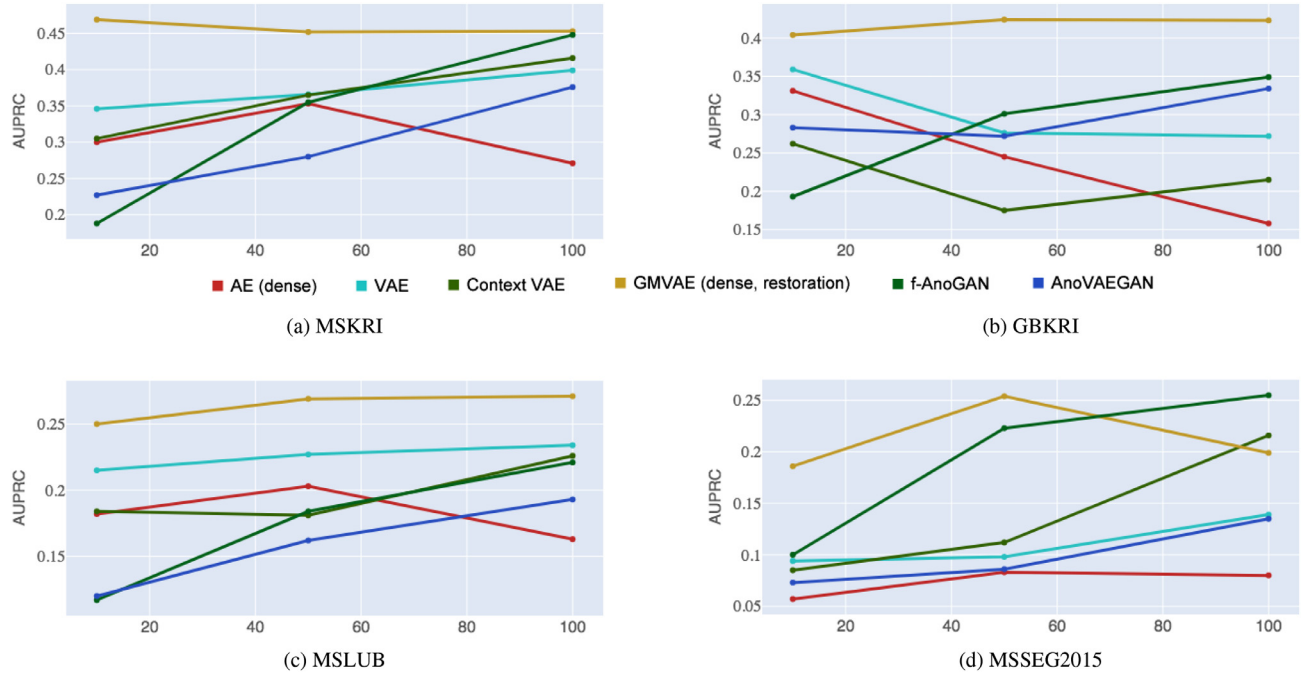


Fig. 8. AUPRC of selected models trained with different numbers of healthy numbers of healthy training subjects (10, 50 and 100%, respectively).

Table 7
Stability analysis of a selection of architecturally and conceptually different approaches (mean and standard-deviation of AUPRC and ⌈DICE⌉ on different datasets).

	$\mathcal{D}_{MS}$		$\mathcal{D}_{GB}$		$\mathcal{D}_{MSLUB}$		$\mathcal{D}_{MSSEG2015}$	
Approach	AUPRC	⌈DICE⌉	AUPRC	⌈DICE⌉	AUPRC	⌈DICE⌉	AUPRC	⌈DICE⌉
AE (dense)	0.260±0.02	0.386±0.01	0.151±0.02	0.293±0.02	0.174±0.01	0.280±0.01	0.079±0.01	0.183±0.01
AE (spatial)	0.167±0.05	0.289±0.05	0.182±0.07	0.303±0.07	0.086±0.03	0.176±0.03	0.055±0.02	0.128±0.04
AE (spatial 8x8x2)	0.353±0.10	0.390±0.03	0.329±0.09	0.417±0.07	0.106±0.02	0.139±0.03	0.181±0.06	0.160±0.01
VAE	0.436±0.01	0.492±0.01	0.293±0.00	0.462±0.00	0.253±0.01	0.347±0.00	0.173±0.03	0.251±0.02
Context AE	0.212±0.02	0.353±0.02	0.273±0.03	0.440±0.02	0.168±0.01	0.268±0.01	0.075±0.01	0.177±0.02
Context VAE	0.380±0.02	0.463±0.02	0.247±0.03	0.377±0.02	0.182±0.02	0.292±0.02	0.091±0.01	0.203±0.01
GMVAE (dense)	0.395±0.00	0.484±0.01	0.338±0.02	0.469±0.01	0.225±0.01	0.301±0.01	0.093±0.00	0.192±0.01
GMVAE (spatial)	0.071±0.03	0.156±0.05	0.130±0.07	0.270±0.06	0.047±0.02	0.139±0.02	0.021±0.01	0.054±0.01
Constrained AE	0.429±0.02	0.469±0.01	0.194±0.03	0.336±0.02	0.217±0.01	0.295±0.00	0.116±0.01	0.251±0.01
Constrained AAE	0.242±0.02	0.365±0.03	0.324±0.02	0.455±0.02	0.193±0.01	0.273±0.02	0.072±0.01	0.194±0.02
f-AnoGAN	0.466±0.02	0.459±0.03	0.332±0.05	0.447±0.05	0.247±0.01	0.316±0.02	0.233±0.00	0.283±0.01
AnoVAEGAN	0.358±0.03	0.432±0.02	0.277±0.01	0.439±0.01	0.222±0.03	0.308±0.02	0.162±0.02	0.266±0.02

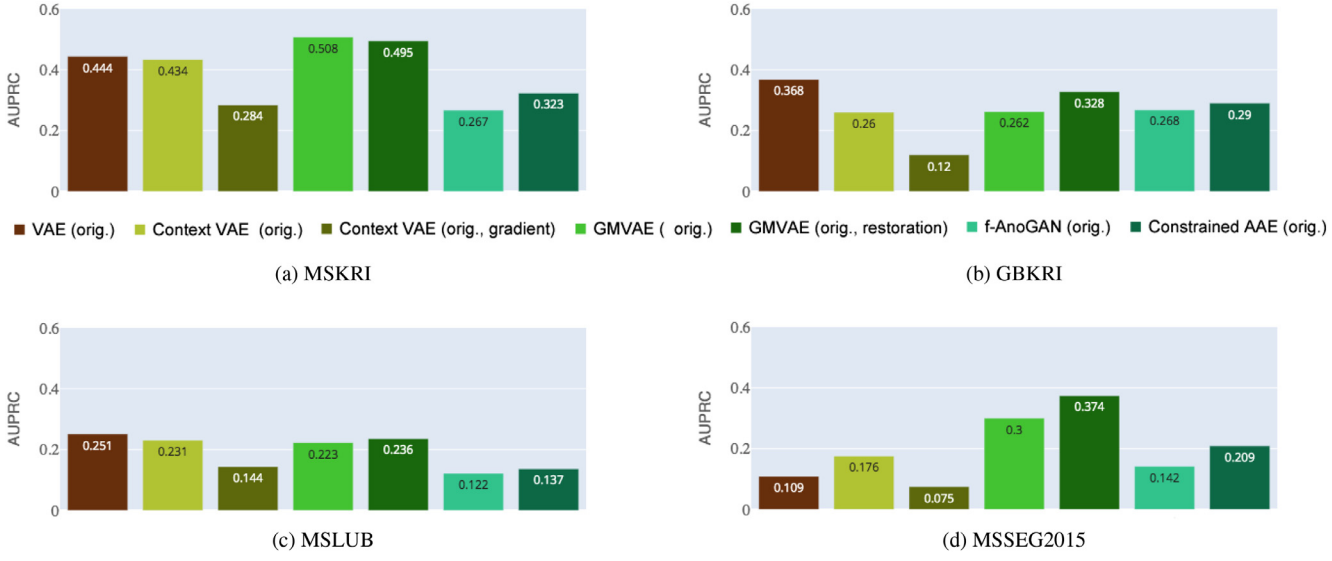
els, which model the healthy distribution the closest due to the Wasserstein-loss, show consistent improvements in AUPRC with a growing training set. Alone the AnoVAEGAN shows a slight drop at 50% of the training data on $\mathcal{D}_{GB}$. The overall top-performer, with one exception, is still the restoration method, here reported using the GMVAE (dense). Alone on $\mathcal{D}_{MSSEG2015}$, this GMVAE shows inconsistent behavior. Both the VAE and Context VAE, our selection from the family of VAEs with a dense bottleneck, show improved and similar performance with increasing number of training subjects on any of the MS datasets. On $\mathcal{D}_{MS}$, both models exhibit inconsistent behavior, and the VAE performs considerably better. Among all the methods, the dense AE yields the most unpredictable performance, varying greatly among different datasets and different number of healthy subjects.

3.15. Stability analysis

To underline that previously reported performances are not just a result of random weight initialization or the stochastic nature of neural network optimization, we further performed a stability analysis on a selection of representative methods using the unified architecture. For this purpose, we trained AEs (dense and spatial), VAEs, GMVAEs (dense and spatial), Context AEs and VAEs, constrained AEs and AAEs, the f-AnoGAN and the AnoVAEGAN. Every method was trained four times in a bootstrapping-inspired manner and tested on all testing data. For every training, out of the 110 healthy training subjects, 100 have been randomly selected for training. Results are reported in terms of the mean and standard deviation of the AUPRC and ⌈DICE⌉ in Table 7. Except for the models with a spatial bottleneck, performances are fairly stable within a standard deviation of 1–3% in both metrics.

3.16. Model complexity

To give some insights on the relation between model complexity, i.e. the number of model parameters and related properties of the network architectures, and segmentation performance, we further rank some of the approaches based on the architectures originally proposed in the respective papers against each other. A comparison is provided on all datasets in Fig. 9. Therein, we find the VAE and the restoration-based GMVAE methods to be stable candidates. Except for $\mathcal{D}_{MSSEG2015}$, the standard VAE approach as proposed in (Baur et al. (2018); Zimmerer et al. (2018, 2019)) shows reliable performance. Similarly, the GMVAE, especially in combination with restoration-based UAD, shows good performance across all datasets. Interestingly, the more complex VAE and Context VAE models in Fig. 9 show only comparable performance to the less complex models following our unified architecture (Fig. 5d. On $\mathcal{D}_{GB}$, none of the more complex models beat the top-performing unified restoration approach. The gradient-based



(a) MSKRI (b) GBKRI

(c) MSLUB (d) MSSEG2015

Fig. 9. AUPRC of all models and UAD approaches, using the original, more complex architectures proposed in the respective papers.

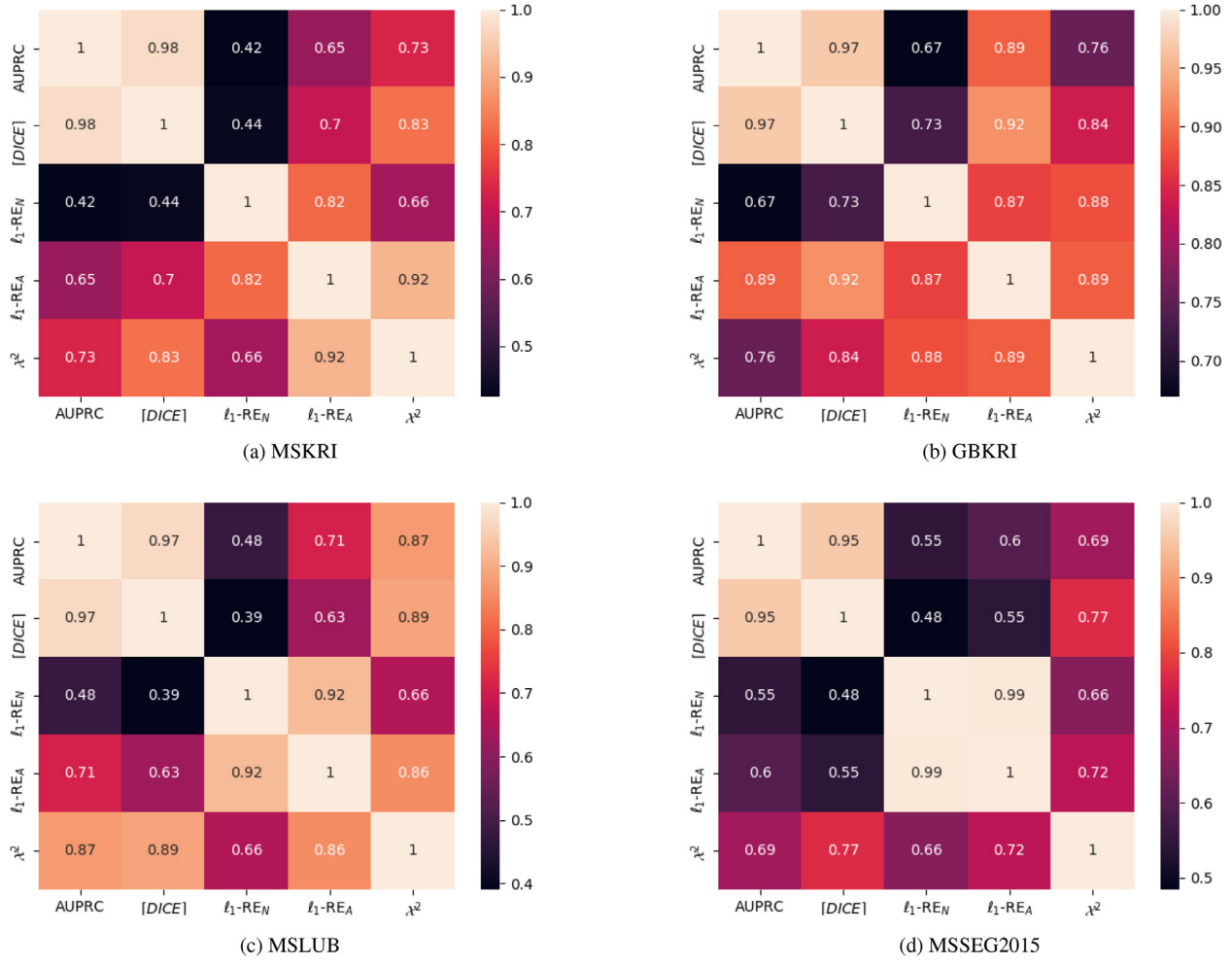


Fig. 10. Correlation matrices among segmentation performance, reconstruction fidelity and overlap among residual histograms of normal and anomalous intensities.

approach, proposed in combination with the original Context VAE, yields lower AUPRC than its unified counterpart. We relate this observation to the reconstruction capabilities of models, which improve with an increase of model parameters. With increasing complexity, larger lesions such as Glioblastoma get reconstructed better as well, which is not desirable.

3.17. Reconstruction fidelity and UAD performance

From Fig. 6 it is clear that apart from spatial models, none of the approaches can reconstruct input perfectly, i.e. none of these methods leave healthy regions intact and substitute anomalous regions with plausible healthy anatomy. Nonetheless, some works perform better than others. We try to relate anomaly segmentation performance to the overlap between a models' residual histograms of normal and anomalous pixels and general reconstruction fidelity. Therefor, we correlate the AUPRC and ⌈DICE⌉ to the $\mathcal{X}^2$-distance of the aforementioned histograms, and further determine how the $\mathcal{X}^2$-distance correlates with reconstruction fidelity of normal and/or anomalous tissue. We do this for every dataset separately to find out if the correlation differs across datasets and pathologies. Fig. 10 shows the correlation heatmaps of aforementioned measures on all datasets.

On $\mathcal{D}_{MS}$ and $\mathcal{D}_{MSLUB}$, AUPRC and ⌈DICE⌉ show moderate to strong correlation to the reconstruction error on anomalous pixels ℓ_1-RE$_A$, but not so much to residuals of normal intensities ℓ_1-RE$_N$. Their correlation to the $\mathcal{X}^2$-distance among residual histograms is the strongest. There is also a strong correlation between $\mathcal{X}^2$ and ℓ_1-RE$_A$, the correlation to ℓ_1-RE$_N$ is less pronounced. From these results we deduce that actual reconstruction fidelity is less important for UAD than clearly distinguishable residual histograms of normal and anomalous intensities.

For $\mathcal{D}_{GB}$, similar, but generally stronger correlations can be seen. Interestingly, there is also a moderate to strong, positive relationship between segmentation performance and magnitude of normal residuals. This indicates that with increasing reconstruction error on both normal and anomalous intensities, segmentation performance improves. We hypothesize that models which reconstruct data well, also reconstruct tumors well. Models with generally poor reconstruction capabilities substitute tumors with poor reconstructions of healthy tissue, leading to better separability between anomalies and normal intensities.

On $\mathcal{D}_{MSSEG2015}$, the previously noticed correlations are hardly present. Instead, ℓ_1-RE$_N$ and ℓ_1-RE$_A$ are strongly correlated and seem to correlate similarly with all other metrics. This clearly reflects the poor contrast in the underlying MR images, which renders UAD unsuitable.

3.18. On the impact of post-processing

|Post-processing of the residuals is vital to obtain viable segmentations, since the compared methods yield residuals where notions of geometry and structure are not explicitly considered. Our chosen post-processing steps introduce very basic prior knowledge

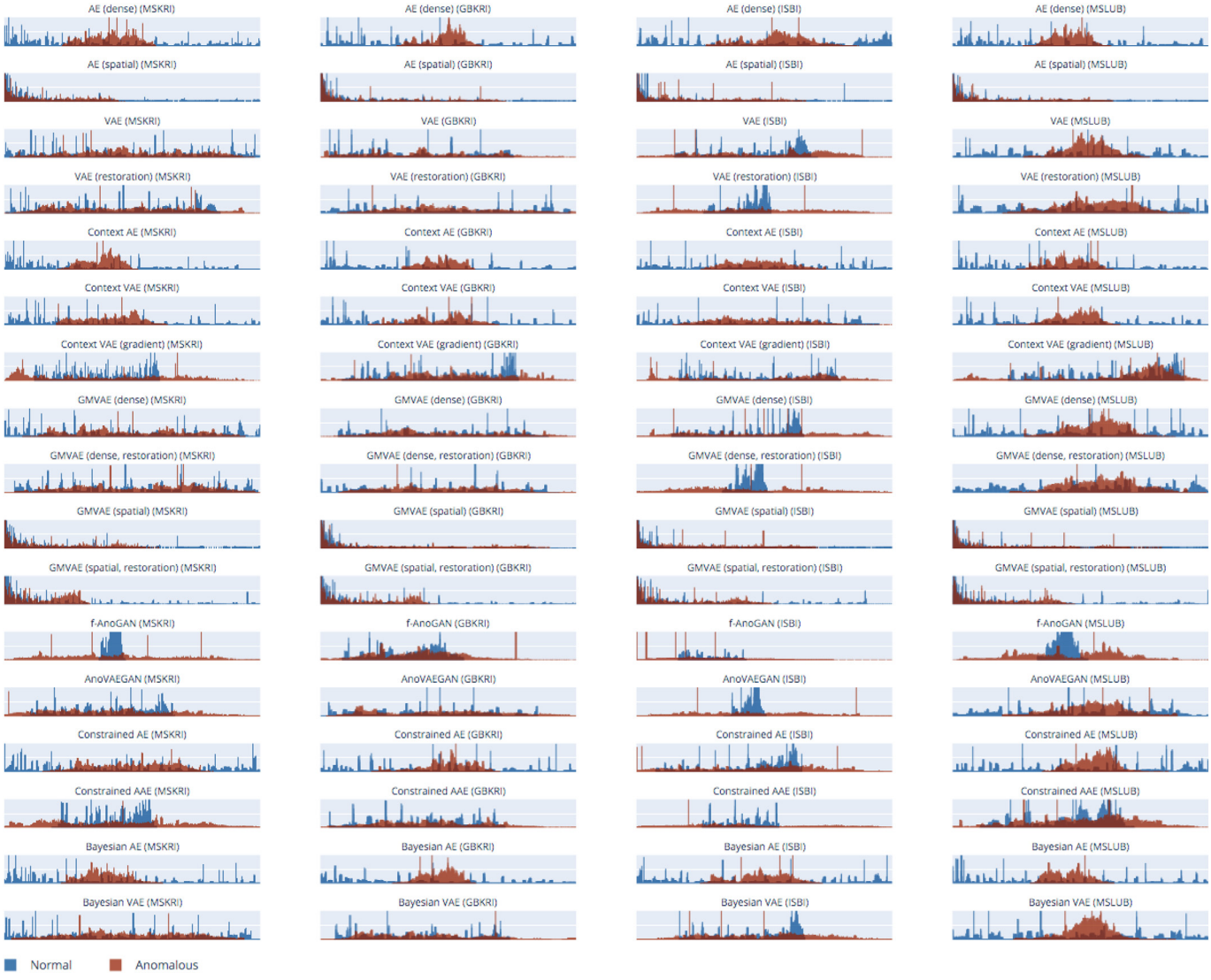


Fig. 11. Normalized histograms of residuals of normal (blue) and anomalous (red) pixels in the intensity range $\in]0; 1.0]$ (ignoring residuals which are completely 0). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

on size, shape and location of anomalies into the inference process and are not designed to favor any method. Yet, they lead to drastic improvements across the board. From Table 8 it can be seen that not using any post-processing leads to very poor segmentation performance across all methods. Discarding very small structures with less than 2 voxels in any direction—by using a mix of median-filtering and connected component analysis—boosts the segmentation performance significantly (see Tables 3, 4, 5 and 6). Inarguably, the incorporation of prior knowledge in the form of post-processing may affect the unsupervised nature of the methods, which leaves room for a philosophical debate.

3.19. Discussion

Ranking|The clear winner of this comparative study is the restoration method applied to a VAE (VAE (restoration)), which achieves best performance on $\mathcal{D}_{MS}$ and $\mathcal{D}_{GB}$, i.e. works best on different pathologies, but also achieves best performance on $\mathcal{D}_{MSLUB}$, i.e. under domain shift. However, there is a downside to the restoration method, namely runtime. A restoration of a single axial slice in 500 iterations takes multiple seconds, which for an entire MR volume accumulates quickly to multiple minutes. The feed-forward nature of purely reconstruction-based approaches allows for a much faster inference. In this context, a very promising method is the reconstruction-based f-AnoGAN, which achieves best performance on the very challenging MSSEG2015 Dataset, and is only slightly inferior to the winning restoration approach on all other datasets. Also, we find that latent variable models perform better in anomaly segmentation than classic AEs. Their reconstructions tend to be more blurry, but the gap between reconstruction errors of normal and anomalous pixels is considerably higher and allows to discriminate much better between anomalies and normal tissue. Among the latent variable models, we find the VAE to be the recommended choice, as it not only performs the best, but is the easiest to optimize. It involves fewer hyperparameters than the other approaches and does not require a discriminator network, which is a critical building block in GANs.

AUPRC vs AUROC|In our experiments, we rely on the AUPRC as a sensible measure for segmentation performance under class imbalance. In contrast, literature primarily reports the AUROC. We want to emphasize once more that we discourage the use of the AUROC, since it favors the more frequent class, i.e. intensities of healthy anatomy. These are, by definition, not only much more frequent in the data, but also reconstructed better. For example, on

Table 8
Results without any post-processing on the different datasets (unified architecture).

	$\mathcal{D}_{MS}$		$\mathcal{D}_{GB}$		$\mathcal{D}_{MSLUB}$		$\mathcal{D}_{MSSEG2015}$	
Approach	AUPRC	⌈DICE⌉	AUPRC	⌈DICE⌉	AUPRC	⌈DICE⌉	AUPRC	⌈DICE⌉
AE (dense)	0.022	0.055	0.045	0.092	0.027	0.065	0.018	0.045
VAE	0.038	0.087	0.092	0.188	0.034	0.082	0.021	0.051
Context AE	0.021	0.053	0.043	0.088	0.026	0.068	0.019	0.051
VAE (Zimmerer)	0.031	0.079	0.079	0.177	0.032	0.083	0.023	0.061
Context VAE (orig.)	0.024	0.063	0.044	0.091	0.027	0.07	0.024	0.06
Context VAE	0.027	0.07	0.052	0.112	0.03	0.078	0.023	0.059
GMVAE (dense)	0.034	0.084	0.081	0.181	0.036	0.09	0.025	0.063
GMVAE (dense restoration)	0.036	0.087	0.086	0.191	0.037	0.093	0.025	0.063
GMVAE (spatial)	0.015	0.041	0.044	0.083	0.021	0.049	0.012	0.028
GMVAE (spatial restoration)	0.014	0.038	0.044	0.082	0.021	0.048	0.012	0.027
f-AnoGAN	0.041	0.103	0.061	0.137	0.034	0.088	0.035	0.089
Constrained AE	0.051	0.125	0.081	0.18	0.044	0.112	0.038	0.097
Constrained AAE	0.036	0.091	0.076	0.173	0.035	0.091	0.027	0.068
VAE (restoration)	0.037	0.089	0.094	0.203	0.038	0.094	0.023	0.058
VAE (restoration w. TV-reg.)	0.036	0.087	0.091	0.199	0.037	0.093	0.023	0.057
AE (spatial)	0.009	0.018	0.037	0.082	0.015	0.033	0.009	0.019
AnoVAEGAN	0.036	0.086	0.057	0.116	0.03	0.07	0.027	0.066

Table 9
Experimenting with different hyperparameters for different models (unified architecture).

	$\mathcal{D}_{MS}$		$\mathcal{D}_{GB}$		$\mathcal{D}_{MSLUB}$		$\mathcal{D}_{MSSEG2015}$	
Approach	AUPRC	⌈DICE⌉	AUPRC	⌈DICE⌉	AUPRC	⌈DICE⌉	AUPRC	⌈DICE⌉
Constrained AE ($\lambda = 0{:}5$)	0.338	0.425	0.349	0.465	0.191	0.28	0.073	0.165
Constrained AAE ($\lambda = 0$)	0.132	0.271	0.23	0.389	0.084	0.2	0.052	0.131
Constrained AAE ($\lambda = 0{:}5$)	0.016	0.041	0.235	0.353	0.022	0.057	0.013	0.025

$\mathcal{D}_{MS}$ the Context VAE (gradient-based method) exhibits the highest AUROC, because it comes with fewer residuals for the more frequent class (healthy pixels), At the same time, it has considerably lower AUPRC than the top-performing VAE (restoration). As a result, the ranking among methods may be inconsistent when comparing them by these different metrics, but AUPRC rates segmentation performance more reliably.

Open Problems|Despite all the recent successes of this paradigm, there are many questions yet to be answered. A key question is how to choose an Operating Point at which the continuous output i) can be binarized and a segmentation can be obtained or ii) an input sample can be considered anomalous. Most of the methods currently either rely on a held-out validation set to determine a threshold for binarization, or make use of heuristics on the intensity distribution. One such heuristic uses the 98th percentile of healthy data as a threshold, above which every value is considered an outlier (Baur et al., 2018). A more elaborate heuristic was used in You et al. (2019), but it relies on False Positives obtained on the healthy training set. It is necessary that more principled approaches for binarization are developed.

Although reconstruction fidelity here is far from perfect, the reviewed methods seem to be indeed capable of segmenting different kinds of anomalies. Nonetheless, we believe that the community should still aim for higher levels of fidelity and modeling MRI also at higher resolution to facilitate segmentation of particularly small brain lesions (e.g. MS lesions, which can become very small) and enhance precision of anomaly localization.

Another obvious downside of the reviewed methods is the necessity of a curated dataset of healthy data. It is debatable whether such methods can actually be called unsupervised or should be seen as weakly-supervised. The community should aim for methods which can be trained from all kinds of samples, even data potentially including anomalies, without the need for human ratings. You et al. (2019) made an initial attempt towards this direction by using a percentile-based heuristic on the training data to mask out potential outliers during training, and with so-called *discriminative reconstruction autoencoders*, (Xia et al., 2015) recently proposed an interesting concept in the Computer Vision field. All in all, more research in this direction is heavily encouraged.

Even though domain shift has not been identified as a burning problem in this comparative study, it is not clear yet to which extent the reviewed approaches are prone to this phenomenon. We encourage a more thorough analysis of it. Alternatively, Federated Learning (McMahan et al., 2017; Rieke et al., 2020) holds the potential to not only preserve data privacy, but to distill domain-specific models into a single model that exhibits improved generalization capabilities.

Generally, the field of Deep Learning based UAD for brain imaging is rapidly growing, and without the availability of a well defined benchmark dataset the field becomes increasingly confusing. This confusion primarily arises from the different datasets used in these works, which come at different resolutions, with different lesion load and different pathologies. All of these properties make it hard to compare methods. Here, we try to give an overview of recent methods, bring them into a shared context and establish comparability among them by leveraging the same data for all approaches. Nonetheless, even the datasets used in this comparative study are limited and many open questions have to remain unanswered. Since UAD methods aim to be general, they need to be evaluated on the most representative dataset possible. Ideally, a benchmark dataset for UAD in brain MRI should comprise a vast number of healthy subjects as well as different pathologies from different scanners, covering the genders and the entire age spectrum.

To date, different works do not only employ different datasets, but also report different metrics. In addition to the benchmark, a clear set of evaluation metrics needs to be defined to facilitate comparability among methods. Very recently, a great leap forward in both directions could be made with a dedicated Medical Out of Distribution (MOOD) analysis challenge (Zimmerer et al., 2020) partially dedicated to UAD in brain MRI, featuring a rich public training dataset of anatomically normal brain MRI data and relying

on thorough analysis & ranking methodology (Wiesenfarth et al., 2019).

Last, the majority of approaches relies on 2D slices, but 3D offers greater opportunity and more context.

4. Conclusion

In summary, we presented a thorough comparison of autoencoder-based methods for anomaly segmentation in brain MRI, which rely on modeling healthy anatomy to detect abnormal structures. We find that none of the models can perfectly reconstruct or restore healthy counterparts of potentially pathological input samples, but different approaches show different discrepancies between reconstruction-error statistics of normal and abnormal tissue, which we identify as the best indicator for good UAD performance.

To facilitate comparability, we relied on a single unified architecture and a single image resolution. The entire code behind this comparative study, including the implementations of all methods, pre-processing and evaluation pipeline is publicly available and we encourage authors to contribute to it. Authors might benefit from a transparent ranking which they can report in their work without having to reinvent the wheel to run extensive comparisons against other approaches.

In our discussion, we also identify different research directions for future work. Comparing different model-complexities, their correlation with reconstruction quality and its effect on anomaly segmentation performance is another research direction orthogonal to our investigations. Determining the correlation between image resolution and UAD performance is also an open task. Further, this study focuses on recent works based on Deep Learning, but a variety on non-Deep-Learning works do exists. A comparison to such "traditional" methods is highly encouraged. However, our main suggestion is the creation of a benchmark dataset for UAD in brain MRI, which involves many challenges by itself, but would be very beneficial to the entire community.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Christoph Baur: Conceptualization, Methodology, Software, Writing - original draft, Formal analysis, Investigation, Visualization, Project administration. **Stefan Denner:** Software, Writing - original draft, Formal analysis, Investigation, Visualization. **Benedikt Wiestler:** Data curation, Resources, Writing - original draft. **Nassir Navab:** Supervision, Resources, Writing - review & editing. **Shadi Albarqouni:** Supervision, Conceptualization, Writing - review & editing, Project administration.

Acknowledgment

The authors would like to thank their clinical partners at Klinikum rechts der Isar, Munich, for generously providing their data. S.A. was supported by the PRIME programme of the German Academic Exchange Service (DAAD) with funds from the German Federal Ministry of Education and Research (BMBF), and B. W. was supported by the DFD SFB-824 grant.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2020.101952

References

Anbeek, P., Vincken, K.L., van Osch, M.J., Bisschops, R.H., van der Grond, J., 2004. Automatic segmentation of different-sized white matter lesions by voxel probability estimation. Med. Image Anal. 8 (3), 205–215.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks. In: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning. PMLR, International Convention Centre, Sydney, Australia, pp. 214–223.

Atlason, H.E., Love, A., Sigurdsson, S., Gudnason, V., Ellingsen, L.M., 2019. Unsupervised brain lesion segmentation from mri using a convolutional autoencoder. In: Medical Imaging 2019: Image Processing, 10949. International Society for Optics and Photonics, p. 109491H.

Baur, C., Wiestler, B., Albarqouni, S., Navab, N., 2018. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. arXiv preprint arXiv:1804.04488.

Bruno, M.A., Walker, E.A., Abujudeh, H.H., 2015. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 35 (6), 1668–1676.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C.H., et al., 2017. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. Neuroimage 148, 77–102.

Chen, M., Kanade, T., Pomerleau, D., Rowley, H.A., 1999. Anomaly detection through registration. Pattern Recognit 32 (1), 113–128.

Chen, X., Konukoglu, E., 2018. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972.

Dilokthanakul, N., Mediano, P.A., Garnelo, M., Lee, M.C., Salimbeni, H., Arulkumaran, K., Shanahan, M., 2016. Deep unsupervised clustering with gaussian mixture variational autoencoders. arXiv preprint arXiv:1611.02648.

Ghosh, P., Sajjadi, M.S., Vergari, A., Black, M., Schölkopf, B., 2019. From variational to deterministic autoencoders. arXiv preprint arXiv:1903.12436.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 2672–2680.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U.G., Kuker, W., Battaglini, M., Rothwell, P.M., et al., 2016. Bianca (brain intensity abnormality classification algorithm): a new tool for automated segmentation of white matter hyperintensities. Neuroimage 141, 191–205.

Iglesias, J.E., Liu, C.-Y., Thompson, P.M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging 30 (9), 1617–1634.

Iheme, L.O., Ünay, D., Baskaya, O., Sennaz, A., Kandemir, M., Yalciner, Z.B., Tepe, M.S., Kahraman, T., Ünal, G.B., 2013. Concordance between computer-based neuroimaging findings and expert assessments in dementia grading.. SIU 1–4.

Jain, S., Sima, D.M., Smeets, D., 2015. Automatic longitudinal multiple sclerosis lesion segmentation: Msmetrix. Longitudinal Multiple Sclerosis Lesion Segmentation Challenge, International Symposium on Biomedical Imaging.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes. In: International Conference on Learning Representations.

Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O., 2015. Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300.

Lesjak, Ž., Galimzianova, A., Koren, A., Lukin, M., Pernuš, F., Likar, B., Špiclin, Ž., 2018. A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. Neuroinformatics 16 (1), 51–63.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., 2016. Adversarial autoencoders. In: International Conference on Learning Representations.

McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. PMLR, pp. 1273–1282.

Menze, B.H., Van Leemput, K., Lashkari, D., Weber, M.-A., Ayache, N., Golland, P., 2010. A generative model for brain tumor segmentation in multi-modal images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 151–159.

Pawlowski, N., Lee, M.C., Rajchl, M., McDonagh, S., Ferrante, E., Kamnitsas, K., Cooke, S., Stevenson, S., Khetani, A., Newman, T., et al., 2018. Unsupervised lesion detection in brain ct using bayesian convolutional autoencoders.

Prastawa, M., Bullitt, E., Ho, S., Gerig, G., 2004. A brain tumor segmentation framework based on outlier detection. Med. Image Anal. 8 (3), 275–283.

Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H., Albarqouni, S., Bakas, S., Galtier, M., Landman, B., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R., Trask, A., Xu, D., Baust, M., Cardoso, M., 2020. The future of digital health with federated learning. npj Digital Medicine 3 (1). doi:10.1038/s41746-020-00323-1.

Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A., 2009. The SRI24 multichannel atlas of normal adult human brain structure. Hum. Brain Mapp. 31 (5), 798–819.

Sato, D., Hanaoka, S., Nomura, Y., Takenaga, T., Miki, S., Yoshikawa, T., Hayashi, N., Abe, O., 2018. A primitive study on unsupervised anomaly detection with an autoencoder in emergency head ct volumes. In: Medical Imaging 2018: Computer-Aided Diagnosis, 10575. International Society for Optics and Photonics, p. 105751P.

Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2019. F-anogan: fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. 54, 30–44.

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker

discovery. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 146–157.
Schmidt, P., 2017. Bayesian inference for structured additive regression models for large-scale problems with applications to medical imaging. lmu.
Sethian, J.A., 1999. Level set methods and fast marching methods: Evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science, 3. Cambridge university press.
Shiee, N., Bazin, P.-L., Ozturk, A., Reich, D.S., Calabresi, P.A., Pham, D.L., 2010. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions.. Neuroimage 49 (2), 1524–1535.
Taboada-Crispi, A., Sahli, H., Hernandez-Pacheco, D., Falcon-Ruiz, A., 2009. Anomaly Detection in Medical Image Analysis. In: Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications. IGI Global, pp. 426–446.
Van Leemput, K., Maes, F., Vandermeulen, D., Colchester, A., Suetens, P., 2001. Automated segmentation of multiple sclerosis lesions by model outlier detection. IEEE Trans Med Imaging 20 (8), 677–688.
Weiss, N., Rueckert, D., Rao, A., 2013. Multiple sclerosis lesion segmentation using dictionary learning and sparse coding.. MICCAI 8149 (Chapter 92), 735–742.
Wiesenfarth, M., Reinke, A., Landman, B.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2019. Methods and open-source toolkit for analyzing and visualizing challenge results. arXiv preprint arXiv:1910.05121.
Xia, Y., Cao, X., Wen, F., Hua, G., Sun, J., 2015. Learning discriminative reconstructions for unsupervised outlier removal. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1511–1519.
You, S., Tezcan, K.C., Chen, X., Konukoglu, E., 2019. Unsupervised lesion detection via image restoration with a normative prior. In: Cardoso, M.J., Feragen, A., Glocker, B., Konukoglu, E., Oguz, I., Unal, G., Vercauteren, T. (Eds.), Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning. PMLR, London, United Kingdom, pp. 540–556.
Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K., 2019. Unsupervised anomaly localization using variational auto-encoders. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 289–297.
Zimmerer, D., Kohl, S.A., Petersen, J., Isensee, F., Maier-Hein, K.H., 2018. Context-encoding variational autoencoder for unsupervised anomaly detection. arXiv preprint arXiv:1812.05941.
Zimmerer, D., Petersen, J., Köhler, G., Jäger, P., Full, P., Roß, T., Adler, T., Reinke, A., Maier-Hein, L., Maier-Hein, K., 2020. Medical out-of-distribution analysis challenge. doi: 10.5281/zenodo.3784230