International Journal of Computer Assisted Radiology and Surgery (2021) 16:1413–1423
https://doi.org/10.1007/s11548-021-02451-9

ORIGINAL ARTICLE

Three-dimensional deep learning with spatial erasing for unsupervised anomaly segmentation in brain MRI

Marcel Bengs[1] · Finn Behrendt[1] · Julia Krüger[2] · Roland Opfer[2] · Alexander Schlaefer[1]

Received: 13 January 2021 / Accepted: 30 June 2021 / Published online: 12 July 2021

Abstract

Purpose Brain Magnetic Resonance Images (MRIs) are essential for the diagnosis of neurological diseases. Recently, deep learning methods for unsupervised anomaly detection (UAD) have been proposed for the analysis of brain MRI. These methods rely on healthy brain MRIs and eliminate the requirement of pixel-wise annotated data compared to supervised deep learning. While a wide range of methods for UAD have been proposed, these methods are mostly 2D and only learn from MRI slices, disregarding that brain lesions are inherently 3D and the spatial context of MRI volumes remains unexploited.

Methods We investigate whether using increased spatial context by using MRI volumes combined with spatial erasing leads to improved unsupervised anomaly segmentation performance compared to learning from slices. We evaluate and compare 2D variational autoencoder (VAE) to their 3D counterpart, propose 3D input erasing, and systemically study the impact of the data set size on the performance.

Results Using two publicly available segmentation data sets for evaluation, 3D VAEs outperform their 2D counterpart, highlighting the advantage of volumetric context. Also, our 3D erasing methods allow for further performance improvements. Our best performing 3D VAE with input erasing leads to an average DICE score of 31.40% compared to 25.76% for the 2D VAE.

Conclusions We propose 3D deep learning methods for UAD in brain MRI combined with 3D erasing and demonstrate that 3D methods clearly outperform their 2D counterpart for anomaly segmentation. Also, our spatial erasing method allows for further performance improvements and reduces the requirement for large data sets.

Keywords Anomaly · Segmentation · Unsupervised · Brain MRI · 3D autoencoder

Introduction

Brain Magnetic Resonance Images (MRIs) allow for three-dimensional (3D) imaging of the brain and are widely used in research and clinical practice for the diagnosis and treatment of neurological diseases. While promising technology advancements of the imaging quality enable an ever-increasing amount of conditions that become detectable [21], reading and interpreting MRI remains a challenging task. First, brain lesion detection and delineation requires expert knowledge and is a tedious time-consuming process, affected by human errors [6]. Second, MRI is increasingly used and hence an ever-increasing amount of images need to be studied, while only a limited number of experts are available [7]. This leads to the urgent need for automatic detection and segmentation of lesions to assist radiologists during clinical practice.

Recently, supervised deep learning methods have shown promising results for this task, while the success of these methods depends heavily on large data sets with high-quality annotations [14]. Note that supervised methods only generalize well to cases that are sufficiently represented in the training data. However, diverse and large annotated data sets are costly to obtain, and often only a few limited cases are available for rare diseases [4].

In contrast to that, human experts can be trained with few healthy cases to generalize, and afterward they are able to detect even arbitrary anomalies without being trained to

Marcel Bengs and Finn Behrendt have contributed equally to this work.

✉ Marcel Bengs
marcel.bengs@tuhh.de

[1] Institute of Medical Technology and Intelligent Systems, Hamburg University of Technology, Hamburg, Germany

[2] jung diagnostics GmbH, Hamburg, Germany

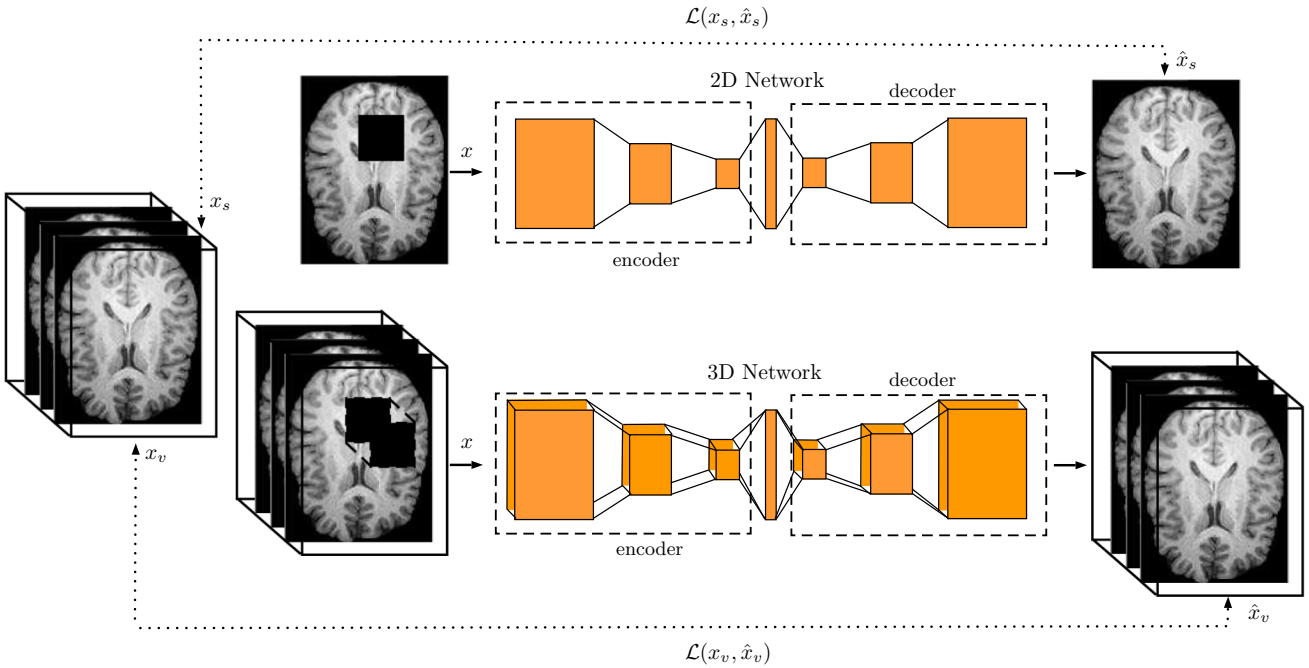

Fig. 1 Our approach for unsupervised anomaly segmentation using 3D deep learning combined with spatial input erasing. For the 2D network, only a single 2D slice x_s is used as input x and volumetric spatial context remains unexploited. Instead, our novel 3D approach receives an entire volume x_v as input x and learns combined features from all spatial dimensions. Also, we propose 3D spatial input erasing, where parts of the input are missing and the network is trained to restore missing image parts. Note, $\hat{x_s}$ and $\hat{x_v}$ refer to the network's reconstruction in 2D and 3D, respectively

an explicit appearance [7]. Deep learning for unsupervised anomaly detection (UAD) follows this concept of identifying unexpected, abnormal data. These methods do not require pixel-level annotations and are only trained with MRI-scans of healthy brains. Here, the task is considered as an anomaly detection problem, where the networks are trained to represent the distribution of healthy anatomy of the human brain and anomalies can be detected as outliers from the learned distribution. Typically, deep learning for UAD follows an encoder–decoder structure trained only on healthy images. Afterward, detection and delineation of pathologies of a test image can be obtained, e.g., by pixel-wise discrepancies between the model's input and reconstruction.

So far, a wide range of deep learning methods have been proposed for UAD in brain MRI, ranging from simple auto-encoders [5] to generative adversarial networks (GANs) [18] focusing on 2D spatial information. These 2D methods have shown promising results; however, the global spatial context provided by MRI volumes remains unused and the inherently 3D structure of brains cannot be learned by the networks. This brings up the question, whether increased spatial context by using entire MRI volumes allows for improved performance, leading to the problem of 3D deep learning for UAD in brain MRI. So far, 3D deep learning for UAD has hardly been considered, only pioneering work in volumetric head CT data has been proposed recently without direct comparison with 2D [17]. 3D deep learning is challenging in nature as it results in an increased representational power that may come with an increased risk of overfitting, leading to poor generalization. For preventing the risk of overfitting, several different regularization strategies have been proposed for deep learning in the context of computer vision. These methods range from simple image transformation such as rotation and flipping to adding noise during the training process, e.g., by stochastically dropping out neuron activations [19] or dropping out entire input regions [9] during training. Especially the latter has been combined with 2D auto-encoder networks, called context-encoders [16], where the networks are enforced to generate the contents of an arbitrary image region conditioned on its surroundings, leading to a better understanding of the global content of the image. This idea has also shown promising results in the context of UAD in brain MRI using 2D methods [22] and might be a promising approach for enforcing the understanding of the global context when entire MRI volumes are used in combination with 3D deep learning.

In this paper, we propose to learn from entire 3D MRI volumes instead of single 2D MRI slices using 3D instead of 2D unsupervised deep learning, shown in Fig. 1. Also, we extend the concept of spatial input erasing for regularization. To this end, we provide an extensive comparison of varia-

tional autoencoders (VAE) with 3D and 2D convolutions and propose several different 3D spatial erasing strategies during training. For our experiments, we use a training data set with brain MRI scans of 2008 healthy patients and evaluate our methods on two publicly available brain segmentation data sets. We focus on T1-weighted MRI data, which are widely used in clinics [1,10], providing a good starting point for anomaly detection. Moreover, we provide an analysis of the impact and the importance of the training data set size, especially in combination with our 3D approach.

Materials and methods

Data set

For training, we consider a data set with anonymized T1-weighted MRI volumes of 2008 healthy subjects from 22 scanners from different vendors. The resolutions in axial direction vary from 0.39mm to 1.25mm with a majority of 1310 samples with 1mm. The slice thickness lies between 0.90mm to 2.40mm with a majority of 906 samples with 1mm. A total of 1506 samples are acquired with a field strength of 1.5 T, 446 samples are acquired with 3 T and 56 with 1 T. Data on all scanners were acquired during clinical routine with a standard 3D gradient echo sequence. All scans were sent to jung diagnostics GmbH for image analysis.

For evaluation, we use two publicly available data sets. First, we consider the publicly available Multimodal Brain Tumor Segmentation Challenge 2019 (BraTS 2019) data set [2,3,15] with T1-weighted image volumes of 335 subjects with the corresponding ground truth segmentation of the tumor. The slice thickness of the BraTS 2019 data set varies from 1mm up to 5mm Second, we use the Anatomical Tracings of Lesions After Stroke (ATLAS) data set [13], which provides T1-weighted image volumes of 304 subjects with corresponding ground truth segmentations of stroke regions. The slice thickness of the ATLAS data set varies from 1mm up to 3mm.

For all image volumes, we apply the following preprocessing. First, we resample all scans to the same isotropic resolution of $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$ using cubic interpolation. Then, we follow the preprocessing of previous studies with 2D deep learning methods for UAD, which include skull stripping, denoising, and standardization [4]. Next, we crop excessive background by using brain masks of the MRI scans and zero-pad all MRI scans to the largest volume resolution in our data set of $191 \times 158 \times 163$. Last, we downsample all volumes to a size of $64 \times 64 \times 64$ for numerical efficiency, as we encounter the computational complexity of 3D deep learning. Regarding our data split for training, we consider 1807 healthy images for training and 201 images for validation of our reconstruction performance. We split our data randomly and stratified by scanners. Considering the images of the BraTS 2019 data set, we randomly sample 133 images for validation and 202 for testing. Using the ATLAS data set, we randomly sample 121 and 183 images for validation and testing, respectively.

Deep learning methods

We address the problem of anomaly segmentation with 2D and 3D unsupervised deep learning methods using 2D MRI slices or 3D MRI volumes, respectively. Given a set of healthy MRI scans, we utilize an encoder–decoder architecture and train our methods to encode to and reconstruct from a lower-dimensional latent space $z \in \mathbb{R}^n$. After the methods are trained, anomalies in a test image can be detected by large reconstruction errors between the input and output image, as the networks are trained to reconstruct only images of healthy brain anatomies, e.g., fail to reconstruct abnormal image areas.

Recently, a comparative study on UAD using 2D deep learning methods [4] has demonstrated that VAE [5,12] allows for promising results, while also being easy to optimize and involving fewer hyperparameters compared to other UAD methods such as GANs. Comparing the VAE with the standard AE, the VAE enforces a structure on the manifold. It has been demonstrated that this leads to performance improvements compared to the standard AE [4]. Hence, we consider the concept of VAEs for our study.

Our general backbone network is shown in Fig. 2 and for the adaption to 2D MRI slices or 3D MRI volumes, we employ 2D or 3D operations for the network, e.g., we use 2D or 3D convolutions. In this way, the architecture details remain the same for 2D and 3D, e.g., the number of layers and feature maps remain same, and only the dimension of the networks operation are changed. Based on our validation set performance, we choose a latent space size of $z \in \mathbb{R}^{128}$ and $z \in \mathbb{R}^{512}$ for our 2D and 3D VAE, respectively.

We study and extend the concept of cutout [9] and context-autoencoders [16], which were proposed for 2D images. The main motivation behind our approach is to further enhance the usage of global image context, especially in combination with 3D methods. Therefore, we propose and evaluate the following different erasing methods for 2D and 3D, which are shown in Fig. 3. Note, we only erase the regions in the input image and not in the ground-truth image that is used for optimization; hence, our networks are enforced to solve an in-painting task for abnormal regions.

First, we simply mask-out a single patch in the input, similar to previous concepts for 2D problems [9,16,22]. Also, we extend this approach to 3D and mask-out a single 3D cube. For the patch and cube erasing method, we randomly select a pixel coordinate within the image as a center point and randomly erase regions with a size from 1% up to 25%

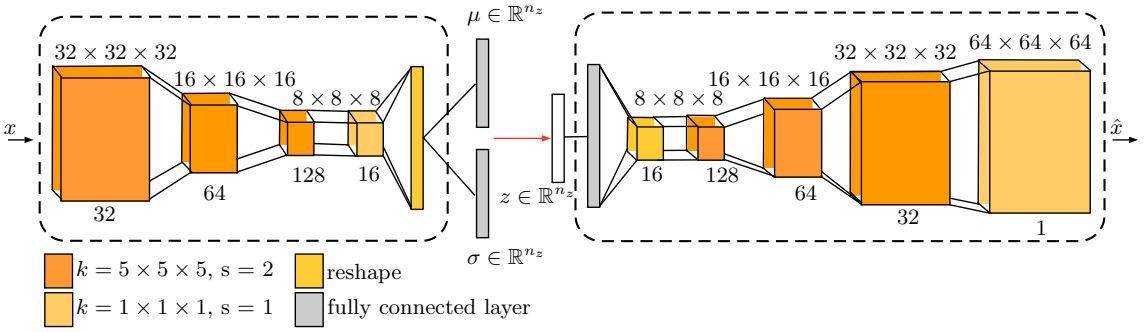

Fig. 2 Our backbone 3D VAD architecture receives input volume $x \in \mathbb{R}^{64\times64\times64}$ and encodes it to the lower-dimensional latent variable $z \in \mathbb{R}^{n_z}$, afterward the decoder reconstructs the output $\hat{x} \in \mathbb{R}^{64\times64\times64}$. The number over the boxes refers to the spatial size; the number below the boxes refers to the number of feature maps. We use convolutions and transposed convolutions in the encoder and decoder, respectively. Note, the first convolution in the encoder downsamples the input from $64 \times 64 \times 64$ to $32 \times 32 \times 32$

of the input size. Note, we refer to this method as patch for 2D and cube for 3D.

Second, we extend this approach and split a single patch or cube into multiple ones. To this end, we mask-out up to ten randomly located and sized patches or cubes within an input image, while the overall erasing size remains in the limit of 1% up to 25% of the input size. We call this method multiple-patch or multiple-cube for 2D and 3D, respectively.

Third, we erase entire brain sides based on the idea of stimulating the networks to exploit the symmetry of a brain. Hence, we randomly erase the right or left side of the brain in the input slice. Similar for 3D, here we randomly erase the right or left side of the brain in 1 up to 32 multiple sequential input slices. We refer to this method as half-slice for 2D and half-volume for 3D.

We systematically evaluate all erasing methods with different strategies for masking-out the regions. First, we simply erase regions in the input, e.g., all intensity values of a region are set to zero similar to previous works [9,16,22]. Second, to further increase the variance of our erasing methods we fill the erased region with noise sampled from the image pixel distribution.

For all our methods, we set the probability of the spatial erasing to $p = 0.5$, such that the network still receives unmodified images.

Training and evaluation

We follow the idea of VAEs; hence, we optimize our networks with respect to the reconstruction loss between the original input image and the network output reconstruction combined with the constraint that the latent variables follow a multivariate normal distribution. Hence, our loss function is based on the l_1-distance between our input and output combined with the distribution-matching Kullback–Leibler divergence for regularization. We train our networks with a batch size of 32 using Adam for optimization with a learning rate of 0.001. We individually tune the number of training epochs of the networks using the reconstruction performance on our validation set with images of healthy subjects.

For all evaluations, we employ the following post-processing steps. First, we multiply each residual image by a slightly eroded brain mask to account for errors occurring at sharp brain-mask boundaries. Next, we remove small outliers with a median filter. For anomaly segmentation of a test image, we consider the voxel-wise residuals obtained from the l_1-distance between the original input image and the network's reconstruction.

For comparison of our methods, we consider voxel-wise anomaly segmentation performance. To this end, we consider the Dice coefficient (DICE) which is defined by

$$\text{DICE} = \frac{2\,|X \cap Y|}{|X| + |Y|}$$

with two sets X and Y. Noteworthy, evaluating the DICE requires binarization of the difference image between the original input image and the network's reconstruction. For this purpose, we utilize our validation set and perform a greedy search to determine the binarization threshold for the segmentation, similar to [4]. Since the scans are normalized, intensity intervals range from 0 to 1. Using the ground truth segmentation, we compute the DICE on the validation set for thresholds at the upper and lower quartile of the center of the intensity interval. Based on the DICE, we cut the interval to either the lower or upper half and continue the search with the updated interval. The procedure is repeated for 10 iterations, and we use the binarization threshold that leads to the best DICE score. Afterward, we use the determined binarization threshold for the test sets. We report the DICE on an entire data set (DICE_D) and also report mean and standard deviation for the subject-wise values (DICE_S). Moreover, to evaluate the models performance for different operating points, e.g.,

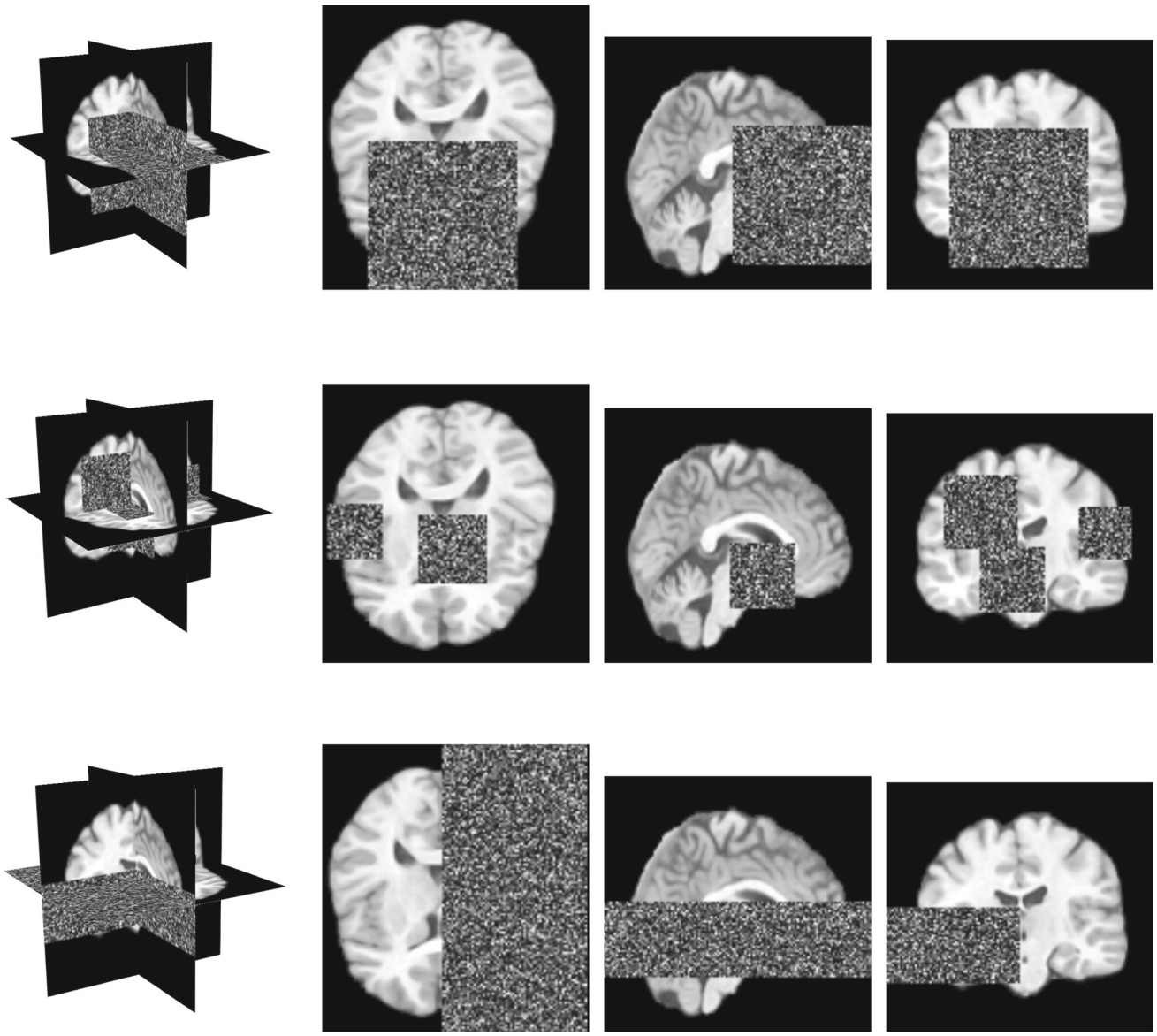

Fig. 3 Our 3D spatial input erasing methods. In each row, sectional planes of a volume with erasing are shown. Top row: We erase a single 3D cube with random location and size (Cube). Middle row: We erase multiple 3D cubes with random location and size (Multi-Cube). Bottom row: We erase an entire brain side in a subvolume (Half-Volume)

binarization threshold for segmentation, we also consider the area under the Precision-Recall-Curve (AUPRC). Here, for each data set, we generate Precision-Recall-Curves (PRC) for each model and then we compute the area under it (AUPRC).

Moreover, we consider our best performing methods and our baseline methods with respect to slice-wise anomaly detection. This allows for localization of anomalies on a slice-level in a volume, i.e., which slice contains a lesion. For this purpose, we divided each volume in our test set into normal and abnormal slices. Considering the lesion annotations, we strictly consider all slices with annotations as abnormal and normal otherwise. For discrimination between normal and abnormal slices, we use the l_1-distance between the original input and the network's reconstruction calculated for each slice. For evaluation of our slice-wise anomaly detection performance independent of the operating point, we report the AUPRC.

Results

First, we compare 2D and 3D UAD deep learning methods combined with our erasing regularization methods in Table 1. For both VAEs, our different erasing methods lead to perfor-

Table 1 Results for our 2D and 3D VAE combined with our spatial erasing methods evaluated on the BraTS 2019 and ATLAS (Stroke) data set

Input and erasing	$DICE_D$	$DICE_S$ ($\mu \pm \sigma$)	AUPRC
BraTS 2019			
2D-None	26.80	25.30 ± 12.37	21.19
3D-None	28.14	26.93 ± 12.40	24.69
2D-Patch-0	27.96	26.52 ± 13.42	22.53
2D-Patch-n	27.99	26.58 ± 13.27	22.54
3D-Cube-0	29.24	27.90 ± 13.57	26.18
3D-Cube-n	30.10	28.80 ± 13.74	27.85
2D-Multi-Patch-0	28.10	26.44 ± 12.89	22.54
2D-Multi-Patch-n	28.51	27.24 ± 13.14	22.81
3D-Multi-Cube-0	28.88	27.67 ± 13.22	25.82
3D-Multi-Cube-n	29.52	28.33 ± 13.42	26.18
2D-Half-Slice-0	26.86	25.44 ± 12.42	21.77
2D-Half-Slice-n	27.97	26.45 ± 13.22	22.84
3D-Half-Volume-0	28.49	27.51 ± 13.17	25.47
3D-Half-Volume-n	28.99	27.92 ± 13.24	26.07
ATLAS (Stroke)			
2D-None	24.72	11.23 ± 13.66	16.86
3D-None	30.68	14.42 ± 16.06	23.74
2D-Patch-0	27.68	12.23 ± 13.67	18.65
2D-Patch-n	27.42	12.36 ± 14.61	18.20
3D-Cube-0	31.50	15.59 ± 17.02	23.47
3D-Cube-n	32.68	15.53 ± 17.30	25.11
2D-Multi-Patch-0	26.99	11.82 ± 14.29	18.72
2D-Multi-Patch-n	28.06	12.88 ± 15.21	19.49
3D-Multi-Cube-0	31.83	15.23 ± 16.64	24.51
3D-Multi-Cube-n	32.37	14.99 ± 17.31	25.13
2D-Half-Slice-0	27.54	11.05 ± 13.70	18.60
2D-Half-Slice-n	28.99	12.13 ± 14.79	20.37
3D-Half-Volume-0	31.00	15.21 ± 17.00	23.14
3D-Half-Volume-n	33.05	15.27 ± 17.21	25.58

The abbreviations for input and erasing refer to the input/VAE dimension, erasing strategy and value used for masking-out a region, e.g., 2D-Patch-0 and 2D-Patch-n stand for a 2D VAE with patch erasing, while the first refers to masking-out a region with zeros and the second refers to masking-out a region with noise

$DICE_D$ represents the metric based on the voxel calculation of an entire data set

$DICE_S$ ($\mu \pm \sigma$) refers to the mean and standard deviation of the subject-wise score

All metrics are in percent

mance improvements. Overall, our 3D VAE outperforms the 2D VAE for all our experiments. Using noise for masking-out the regions works slightly better than masking-out with zeros. For our 3D VAE using a single cube for erasing, followed by masking-out an entire brain side in a subvolume works best. Considering our 2D-VAE, masking-out an entire brain side shows the best results, closely followed by masking-out multiple patches. Comparing the $DICE_D$ of our best performing 3D approach (3D-Cube-n) with the 2D baseline approach (2D-None) demonstrates a relative performance improvement of 12.31% and 32.20% on the BraTS 2019 and ATLAS data set, respectively.

Second, we evaluate the performance of our baselines and best performing methods with respect to lesion size in Fig. 4. Here, our results demonstrate that the smallest and largest lesions are challenging. Consistently, using erasing improves the $DICE_S$ over all lesion sizes, while being particularly effective for large lesions. Also, comparing 2D and 3D methods shows that 3D consistently outperforms 2D, especially for small lesions.

Third, we evaluate the effect of the data set size in Fig. 5. Reducing the data set has a pronounced impact on the performance for 3D as well as 2D, especially when less than 60% of the training data is used. Also, the spatial erasing works better when the network is trained with more data. While reducing the data set size has a larger impact on 3D, even with only 20% of the training data the 3D VAE works better than the 2D VAE with erasing and 100% of the training data. Moreover, our erasing turns out to be effective for the 2D VAE, considering that a 2D VAE without erasing trained with 100% of data is outperformed by a 2D VAE with erasing trained with only 20% of the data.

Fourth, Fig. 7 demonstrates example images for our best performing method 3D-Cube-n. Notably, the ground truth segmentation is highlighted in all difference images, while also showing errors at further regions.

Moreover, we use our best performing 2D and 3D methods trained on T1-weighted MRI data and evaluate on T1ce-weighed MRI data from the BraTS 2019 data set to study the effect of using additional image information, see Table 2. Here, we observe immediate performance improvements compared to T1-weighting for both 2D and 3D with a relative improvement of 13.61% and 21.82% for 2D and 3D considering the $DICE_D$.

Last, we evaluate our baseline and best performing methods with respect to slice-wise anomaly detection, see Fig. 6. Here, our best performing method achieves an AUPRC of 71.2%. Also for this task using 3D information and erasing turns out to be beneficial, improving the AUPRC by approximately 4% compared to the 2D VAE.

Discussion

We consider the problem of unsupervised anomaly segmentation and propose to learn from entire 3D MRI volumes instead of single 2D MRI. For this purpose, we extend 2D VAEs to 3D and also propose several different input erasing methods for regularization. Comparing our 2D VAE (2D-None) with the corresponding 3D version (3D-None) without any input

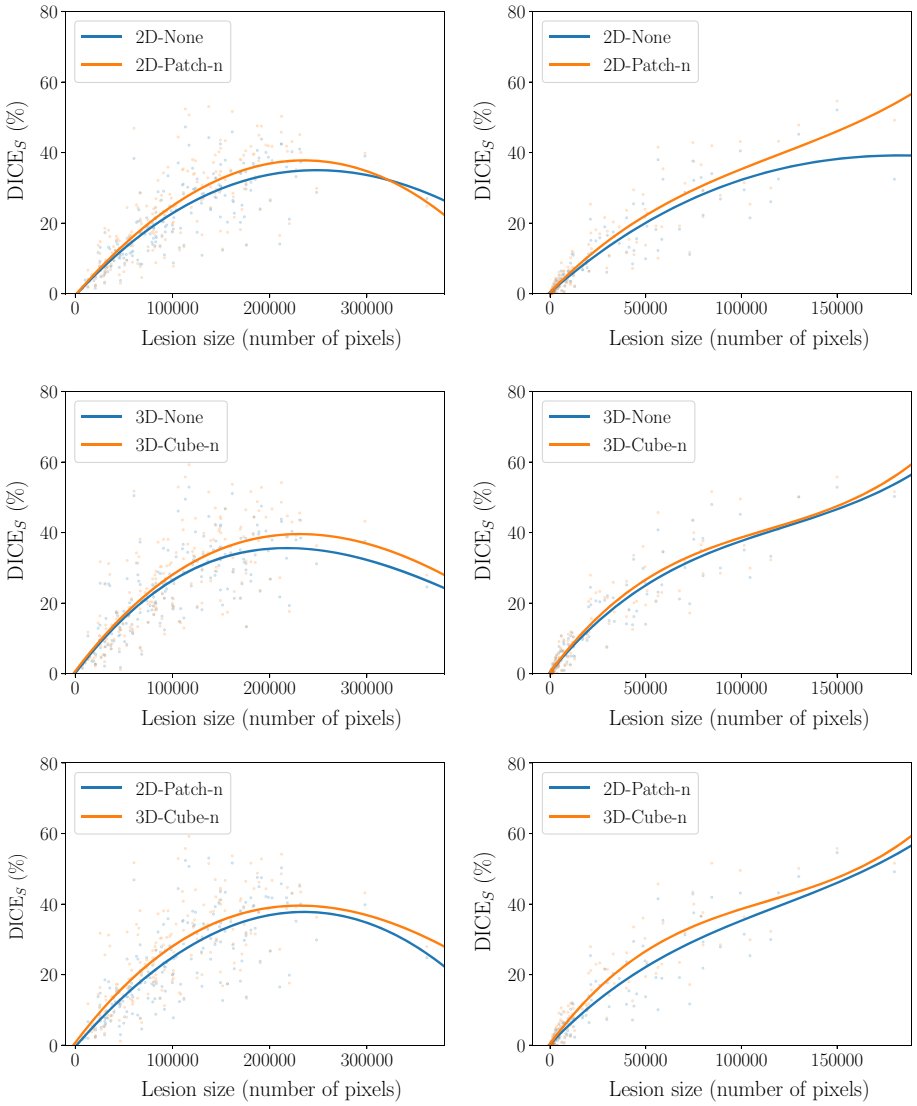

Fig. 4 Subject-wise DICE$_S$ over lesion size. Lesion size refers to the number of annotated pixels for the lesion. Results for the BraTS 2019 data set and ATLAS data set are shown left and right, respectively. (Top) Comparing 2D VAE with and without erasing; (Middle) Comparing 3D VAE with and without erasing; (Bottom) Comparing 2D and 3D VAE with erasing. Transparent dots refer to the subject-wise DICE$_S$ scores. Solid lines are derived by a polynomial regression of order three

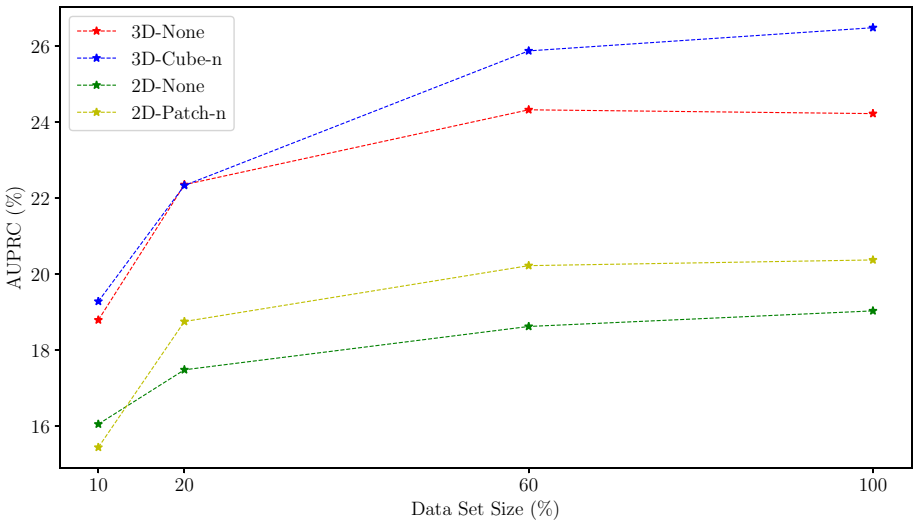

Fig. 5 Impact of data set size on the UAD performance. We train our methods with 10%, 20%, 60%, and 100% of the training data, shown is the average AUPRC using our two test data sets (BraTS 2019, ATLAS)

Table 2 Results for additional image information considering the BraTS 2019 data set

Input and erasing	Sequence	$DICE_D$	$DICE_S$ ($\mu \pm \sigma$)	AUPRC
2D-Patch-n	T1	27.99	26.58 ± 13.27	22.54
2D-Patch-n	T1ce	31.80	29.08 ± 12.77	24.28
3D-Cube-n	T1	30.10	28.80 ± 13.74	27.85
3D-Cube-n	T1ce	36.67	33.40 ± 14.55	31.12

$DICE_D$ represents the metric based on the voxel calculation of an entire data set
$DICE_S$ ($\mu \pm \sigma$) refers to the mean and standard deviation of the subject-wise score
All metrics are in percent

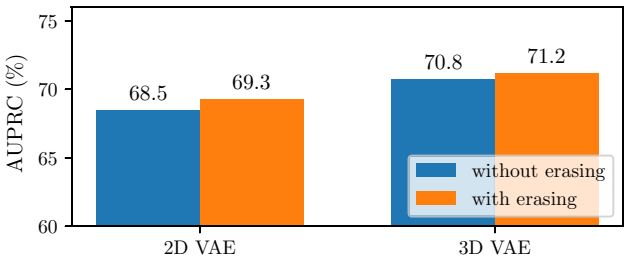

Fig. 6 Slice-wise anomaly detection for our baseline and best performing methods. Shown is the AUPRC on the combination of our test sets (BraTS 2019, ATLAS). 2D VAE with and without erasing refers to 2D-None and 2D-Patch-n, respectively. 3D VAE and without erasing refers to 3D-None and 3D-Cube-n, respectively

erasing demonstrates that 3D outperforms the 2D version on two public data sets, especially for the stroke data set with a $DICE_D$ of 30.68% for 3D compared to a $DICE_D$ of 24.72% for 2D, see Table 1. This highlights that 3D information can be effectively leveraged by a 3D VAE and agrees with our expectation that increased spatial context by using entire MRI volumes allows for improved anomaly segmentation performance.

We also evaluate 2D and 3D input erasing for regularization and train the networks to restore missing image parts conditioned on its surroundings. Our results in Table 1 demonstrate that input erasing allows for further performance improvements both for our 2D and 3D VAE. Regarding the method for masking-out a region, previous works in 2D mostly simply mask our input regions with zeros [9,16,22]. However, our results demonstrate that using noise for masking-out a region in the input works slightly better, indicating that the increased variance during training is advantageous for regularization.

We also consider different strategies such as erasing multiple patches or an entire brain side. While all erasing strategies are beneficial, there is no clear winner between the different strategies considering our results on both data sets. Furthermore, one could argue that our input erasing leads to brain anatomy that deviates from normal, which is in slight contrast to the idea of only providing healthy brain anatomy as input. However, our ground-truth image that is used for optimization remains unmodified; hence, our networks are enforced to solve an in-painting task for abnormal regions. Our results demonstrate that this leads to an improved segmentation performance.

To gain further insights, we study the performance with respect to the lesion size in Fig. 4. While providing consistent performance improvements, erasing turns out be especially valuable for larger lesions. This might be attributed to the fact that with erasing, networks are enforced to solve an additional in-painting task, making them suited to handle inputs with large anomalies. Also, our results in Fig. 4 further emphasize the value of 3D information, especially for smaller lesions considering the ATLAS data set.

Next, we study the effect of the training data set size. As expected, the data set size has a notable impact on the performance, see Fig. 5. It stands out that our 3D methods trained with only 20% of the training data even outperform the 2D methods trained with 100% of the data. This indicates that increasing the spatial context during training is even more important than increasing the data set size. This is an interesting observation, as one could assume that due to the increased number of parameters, 3D-Models require more data compared to their 2D-counterparts. We believe that this counter-intuitive behavior could explained by the increased complexity of the task and the bigger input image for the 3D approach. The learning task of the 3D model can be considered more complex since an entire volume must be processed and reconstructed at once, while 2D is only trained to process a single slice. Also, for 3D the input image is bigger (volume) compared to 2D (single slice). Note, if the input image is bigger, then a network might need more expressive power to capture the patterns in the input image, as shown in [20].

Considering our erasing approach and the data set size suggests that solving the additional in-painting task needs sufficient training data to provide effective regularization. However, with only 60% of the training data our models with our regularization approach lead to higher performance than a model without regularization trained with the full dataset. We argue this demonstrates the effectiveness of our regularization approach, as less data are required to achieve similar or better performance compared to a model without regularization. Still, increasing the data set size is valuable as the

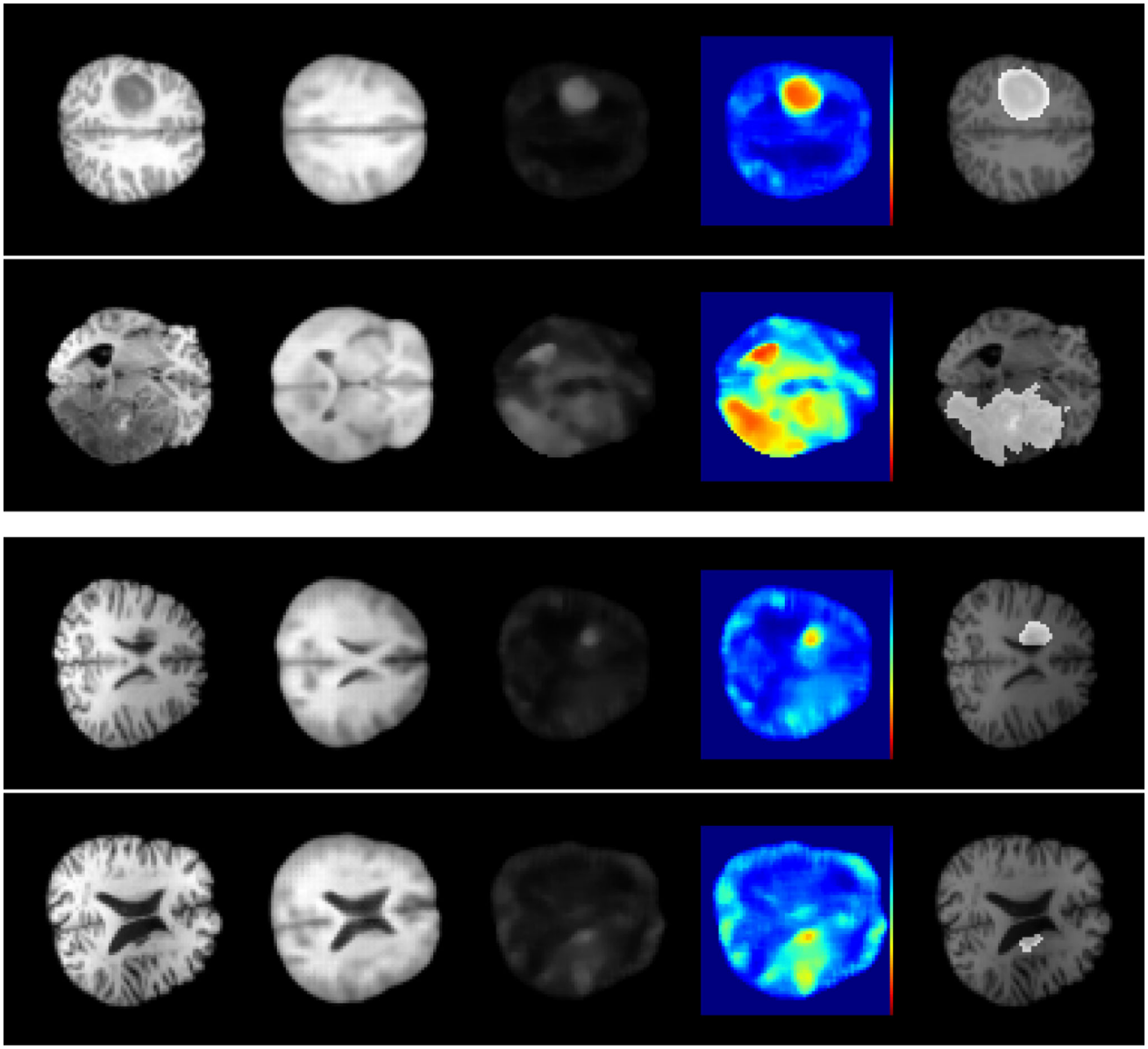

Fig. 7 Four example test cases using our best performing method 3D-Cube-n. From left to right: Input image, output image, difference image, heat-map difference image, and ground truth segmentation. The first two lines contain examples from the BraTS 2019 data set and the two bottom lines contain examples from the ATLAS data set

performance for our model with erasing continues to improve with a larger training data set.

Comparing our novel 3D methods with input erasing with the previous 2D approach demonstrates a relative performance improvement of 12.31% and 32.20% on the BraTS 2019 and ATLAS data set, respectively. A comparable work evaluating UAD performance on the same ATLAS data set achieves a mean subject-wise DICE score of $12 \pm 12\%$ with their best performing method [8]. Notably, this 2D method is restoration-based and involves significantly increased computational complexity. Our 3D approach with input erasing leads to a mean subject-wise DICE score of $15.53 \pm 17.30\%$, improving the UAD state-of-the-art on this data set. This demonstrates the effectiveness of our approach. Comparing our results on the BraTS 2019 data set with other works that utilize additional image information, e.g., T2-weighted data [8,22], highlights the advantage of additional image information. Similar, we observe immediate performance improvement for our methods when evaluated on T1ce-weighted data, despite the domain adaption from T1, see Table 2. Also, other studies that use multiple MRI sequences [4,5] achieve higher performance metrics; however, a direct

comparison is difficult due to different data sets and settings. Notably, multiple MRI sequences are beneficial but not always available [1,10], imposing an additional challenge on UAD.

Putting UAD into perspective with supervised methods demonstrates that segmentation performance is in a moderate range. Considering the BRATS 2019 data set, supervised methods achieve a mean subject-wise DICE score of around 90% [11] utilizing all available MRI sequences (T1, T1ce, T2, FLAIR). Considering the ATLAS data set, supervised methods achieve mean subject-wise DICE scores in the range of 32.92% up to 53.49% [10]. While UAD is notably more challenging than supervised segmentation, the overall UAD performance on these supervised data sets might also be limited, as the annotation focuses on pre-specified lesions and not all anomalies in the images might be labeled. This is also demonstrated in Fig. 7, where, e.g., the segmentation focuses only on the tumor and not on all brain regions that deviate from normal. Also, the domain shifts between different data sets might be challenging, which is also pointed out in previous works [4,22].

Considering these challenges, we also evaluate our methods with respect to slice-wise anomaly detection, see Fig. 6. Here, we observe significantly increased performance compared to segmentation with an AUPRC of 71.2% for our best performing method. The slice-wise detection performance motivates that UAD can be helpful in red-flagging suspicious MRI data in clinical routine, especially with T1-weighted MRI data. Also, we believe that unsupervised segmentation gives additional cues to the reader as to where an anomaly may be located and thus, it is helpful to quickly localize a potential anomaly or lesion. For this, our work consists a valuable contribution by demonstrating the benefits and emphasizing the use of 3D-models with spatial erasing for voxel-wise and slice-wise UAD.

For future work, our findings could be extended to more complex deep learning methods for UAD, such as GANs [18]. In particular, combining our 3D approach with restoration-based methods [8] might improve the overall performance. However, this approach also leads to significantly increased runtime and computational efforts, e.g., a restoration accumulates quickly to multiple minutes for a single MRI [4], which is particularly challenging for clinical routine.

Conclusion

We study the task of unsupervised anomaly segmentation in brain MRI and propose to use entire 3D MRI volumes instead of single 2D MRI slices by extending 2D VAEs to 3D. Also, we study and extend the concept of input erasing and propose several different 3D input erasing strategies for regularization. Overall, our results demonstrate that using increased spatial context by using entire MRI volumes combined with 3D deep learning clearly outperforms 2D methods. Also, we observe that combining deep learning with spatial input erasing allows for further performance improvements and reduces the requirement for large training data sets.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was partially funded by Grant Number ZF4026303TS9.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This work was conducted retrospectively on data from clinical routine which was completely anonymized. Ethical approval was therefore not required. Also this work relies on the BraTS 2019 and ATLAS data set. For use of these data sets, no ethics statements are necessary.

References

1. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ (2017) Deep learning for brain MRI segmentation: state of the art and future directions. J Digit Imaging 30(4):449–459
2. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, Freymann JB, Farahani K, Davatzikos C (2017) Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data 4:170117
3. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A et al (2018) Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629
4. Baur C, Denner S, Wiestler B, Navab N, Albarqouni S (2021) Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. Med Image Anal 101952
5. Baur C, Wiestler B, Albarqouni S, Navab N (2018) Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In: International MICCAI brainlesion workshop, pp 161–169. Springer
6. Bruno MA, Walker EA, Abujudeh HH (2015) Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 35(6):1668–1676

7. Chen X, Konukoglu E (2018) Unsupervised detection of lesions in brain MRI using constrained adversarial auto-encoders. In: International conference on medical imaging with deep learning
8. Chen X, You S, Tezcan KC, Konukoglu E (2020) Unsupervised lesion detection via image restoration with a normative prior. Med Image Anal 64:101713
9. De Vries T, Taylor GW (2017) Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552
10. Ito KL, Kim H, Liew SL (2019) A comparison of automated lesion segmentation approaches for chronic stroke T1-weighted MRI data. Hum Brain Mapp 40(16):4669–4685
11. Jiang Z, Ding C, Liu M, Tao D (2019) Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In: International MICCAI brainlesion workshop, pp 231–241. Springer
12. Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114
13. Liew SL, Anglin JM, Banks NW, Sondag M, Ito KL, Kim H, Chan J, Ito J, Jung C, Khoshab N, Lefebvre S, Nakamura W, Saldana D, Schmiesing A, Tran C, Vo D, Ard T, Heydari P, Kim B, Aziz-Zadeh L, Cramer S, Liu J, Soekadar S, Nordvik JE, Westlye L, Wang J, Winstein C, Yu C, Ai L, Koo B, Craddock R, Milham M, Lakich M, Pienta A, Stroud A (2018) A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. Sci Data 5:180011
14. Lundervold AS, Lundervold A (2019) An overview of deep learning in medical imaging focusing on MRI. Z Med Phys 29(2):102–127
15. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahaniy K, Kirby J, Burren Y, Porz N, Slotboomy J, Wiest R, Lanczi y L, Gerstnery E, Webery MA, Arbel T, Avants B, Ayache N, Buendia P, Collins L, Cordier N, Van Leemput K et al (2015) The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 34(10):1993–2024
16. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA (2016) Context encoders: feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2536–2544
17. 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, vol 10575, p 105751P. International Society for Optics and Photonics
18. Schlegl T, Seeböck P, Waldstein SM, Langs G, Schmidt-Erfurth U (2019) f-AnoGAN: fast unsupervised anomaly detection with generative adversarial networks. Med Image Anal 54:30–44
19. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
20. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning, pp 6105–6114. PMLR
21. Vernooij MW, Ikram MA, Tanghe HL, Vincent AJ, Hofman A, Krestin GP, Niessen WJ, Breteler MM, van der Lugt A (2007) Incidental findings on brain MRI in the general population. N Engl J Med 357(18):1821–1828
22. Zimmerer D, Kohl SA, Petersen J, Isensee F, Maier-Hein KH (2019) Context-encoding variational autoencoder for unsupervised anomaly detection. In: International conference on medical imaging with deep learning