Computers in Biology and Medicine 154 (2023) 106610

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

# Unsupervised anomaly detection in brain MRI: Learning abstract distribution from massive healthy brains

Guoting Luo [a,1], Wei Xie [b,1], Ronghui Gao [b], Tao Zheng [c], Lei Chen [d,**], Huaiqiang Sun [a,*]

[a] Huaxi MR Research Center (HMRRC), Department of Radiology, West China Hospital of Sichuan University, Chengdu, China
[b] Department of Radiology, West China Hospital of Sichuan University, Chengdu, China
[c] IT Center, West China Hospital of Sichuan University, Chengdu, China
[d] Department of Neurology, West China Hospital of Sichuan University, Chengdu, China

ARTICLE INFO

*Keywords:*
Anomaly detection
Unsupervised learning
Autoencoder
Brain MRI

ABSTRACT

*Purpose:* To develop a general unsupervised anomaly detection method based only on MR images of normal brains to automatically detect various brain abnormalities.
*Materials and methods:* In this study, a novel method based on three-dimensional deep autoencoder network is proposed to automatically detect and segment various brain abnormalities without being trained on any abnormal samples. A total of 578 normal T2w MR volumes without obvious abnormalities were used for model training and validation. The proposed 3D autoencoder was evaluated on two different datasets (BraTs dataset and in-house dataset) containing T2w volumes from patients with glioblastoma, multiple sclerosis and cerebral infarction. Lesions detection and segmentation performance were reported as AUC, precision-recall curve, sensitivity, and Dice score.
*Results:* In anomaly detection, AUCs for three typical lesions were as follows: glioblastoma, 0.844; multiple sclerosis, 0.858; cerebral infarction, 0.807. In anomaly segmentation, the mean Dice for glioblastomas was 0.462. The proposed network also has the ability to generate an anomaly heatmap for visualization purpose.
*Conclusion:* Our proposed method was able to automatically detect various brain anomalies such as glioblastoma, multiple sclerosis, and cerebral infarction. This work suggests that unsupervised anomaly detection is a powerful approach to detect arbitrary brain abnormalities without labeled samples. It has the potential to support diagnostic workflow in radiology as an automated tool for computer-aided image analysis.

## 1. Introduction

The common anomaly in brain include glioblastomas, multiple sclerosis (MS), cerebral infarction (CI) and so forth. Gliomas are the most common primary central nervous system tumor, accounting for almost 50% of patients with primary intracranial tumors, which can be classified into low-grade (LGG) and high-grade (HGG) types based on their malignancy [1,2]; Multiple sclerosis (MS) is a chronic, immune-mediated, demyelinating disorder of the central nervous system, which must be diagnosed early and accurately because it is still a major cause of neurological disability in young adults; Cerebral infarction is ischemic necrosis of brain tissue caused by a blood clot or embolus blocking a blood vessel, which remains one of the leading causes of morbidity and mortality worldwide [3–5].

Magnetic resonance imaging (MRI) is an essential modality to detect brain abnormalities as it is able to provides various tissue contrasts from planes of arbitrary directions, showing variations as changing in size, shape, location, and intensity [6,7]. MRI produces images with unmatched soft-tissue contrast and spatial resolution. The mechanism of MRI to produce tissue contrast is to manipulate proton spins through radio pulse sequence, causing the protons in different tissues to produce signals of different intensities. For example, protons in free water can generate higher signal than those in protein or lipid under the T2-weighted (T2w) sequence. Necrosis caused by tumor growth or edema caused by inflammation will increase the content of free water in the tissue, making these lesions appear hyperintense on T2-weighted

* Corresponding author.
** Corresponding author.
*E-mail addresses:* leilei_25@126.com (L. Chen), sunhuaiqiang@scu.edu.cn (H. Sun).
[1] Guoting Luo and Wei Xie contributed equally to this work.

https://doi.org/10.1016/j.compbiomed.2023.106610
Received 5 August 2022; Received in revised form 20 January 2023; Accepted 22 January 2023
Available online 25 January 2023

(T2w) images. Thus, abnormally high signal on T2-weighted images is an important sign for diseases diagnosis.

To assist radiologists in improving workflow efficiency and diagnostic accuracy using MRI images, numerous algorithms have been proposed to automatically detect abnormalities. Among them, supervised deep learning-based methods have achieved state-of-the-art performance. In the early stages, segmentation algorithms based on partial differential equation, such as level sets, were used to segment tumors from 3D volumes for volume evaluation [8]. As well as graph cuts [9] proved to be a powerful interactive segmentation technique for semi-automatic segmentation of multiple sclerosis lesions in MRI. In recent years, deep learning-based methods [10–12] have been proposed to effectively detect and segment medical images. For example, numerous brain MRI image segmentation algorithms have been proposed in public challenges such as Multimodal Brain Tumor Image Segmentation (BraTs) [13] and Ischemic Stroke Lesion Segmentation (ISLES) [14] categories for anomaly detection and segmentation. As well as gastrointestinal tract lesions segmentation and COVID-19 X-ray image segmentation [15–18]. However, such supervised approaches require vast amounts of annotated data, which are scarce and costly to obtain. Another common drawback of those supervised methods is their 'point solution' design, which means that deep neural networks designed for a specific disease perform poorly on unseen sorts of anomalies.

Given these constraints, unsupervised anomaly detection (UAD) has been emerging as an attractive alternative approach. UAD typically models only the distribution of the healthy tissue, which allows the model to reconstruct only healthy brain anatomy, but not regions of abnormality. Regions with large differences between input and output suggest the presence of abnormalities. Therefore, it is able to take full advantage of large datasets from healthy samples, avoiding the costly manual labeling of abnormal samples and the subsequent issues involved in training with highly class-imbalanced data [19].

Furthermore, it theoretically allows to detect arbitrary, even rare pathologies which supervised approaches might fail to find. In the early stage of UAD research, UAD was considered as the one-class classification problem, such as one-class SVM [20] and deep one-class networks [21]. Unsupervised clustering methods, such as the k-means method and Gaussian Mixture Models (GMM) [22,23], have also been applied to build a detailed profile of the normal data for identifying the anomalies. However, these methods usually suffer from suboptimal performance when processing high-dimensional data [24]. Reconstruction-based methods are proposed relying on an assumption that the anomalies cannot be represented and reconstructed accurately by a model learned only on normal data [25]. Some very recent works trained deep AutoEncoders (AEs) for anomaly detection. And deep generative models such as Variational AutoEncoders (VAEs) and generative adversarial networks (GANs) were also used to anomaly detection [26,27]. To be specific, Schlegl et al. [28] have developed a generative adversarial network (f-AnoGAN) for anomaly detection on optical coherence tomography images. Tian et al. [19] used constrained contrastive distribution learning for anomaly detection on three different colonoscopy and fundus screening datasets. For brain MRI, Hespen et al. [29] used a conditional generative adversarial network (GANomaly) to identify chronic brain infarcts. Lambert et al. [30] proposed VAE architecture for unsupervised segmentation of various diseases of the brain. Baur et al. [31] did a comparative study of common deep learning-based UAD approaches on three multiple sclerosis datasets. In another recent work [32], Baur et al. leveraged spatial AE with skip-connections for brain lesion detection and segmentation.

However, most researches related to UAD decomposed MRI volumes into a collection of independent 2D slices [29,31–33]. Despite the reduced computational difficulty, however, 2D slices that contain no or little brain tissue may have the potential to adversely affect model training. Secondly, most lesions exist in continuous slices where contextual information is essential for differential diagnosis. Using two-dimensional data would result in the loss of this important spatial information. Another limitation of these methods is that they are limited to a monocentric dataset during training and testing, which fails to evaluate the generalization performance of generated model.

To this end, we proposed a 3D residual autoencoder framework for UAD on whole-brain MR images. The proposed model learns to reconstruct the healthy brain volumes with relative high spatial resolution. The segmentation of abnormalities are generated by the difference between the reconstructed volumes and the input volumes with abnormalities.

The contributions of the current study are highlighted as follows.

1) An unsupervised deep learning framework was proposed to detect various brain abnormalities in MR images without the need for lesion labeling.
2) Allows for slice-level anomaly identification and voxel-level segmentation.
3) The deep neural network, which was trained on only massive healthy brains, exhibited excellent testing generalization performance across one public data and two in-house data.

## 2. Materials and Methods

### 2.1. Patient datasets

This retrospective study was approved by the local institutional ethics review board, and informed consent was waived. In this work, three different brain MR datasets were utilized: IXI dataset (all normal brains) for model training, BraTs dataset (all glioblastomas) and two in-house datasets (containing multiple sclerosis and cerebral infarction) for model testing. It's worth noting that the MRI sequences provided by IXI and BraTs are not the same, but both contain T2-weighted (T2w) images. Therefore, only the interaction of these two datasets, T2w images, was included in this study. One limitation of deep learning is the fragile general relevance, which is shown by the fact that architectures trained on one database with a unique setting can hardly achieve great performances in different datasets. To demonstrate the general applicability of the proposed method, we trained the model using only the IXI dataset and tested it on the other two datasets. Details of all datasets are listed in Table 1.

**IXI dataset:** This dataset contains MR images of 578 healthy subjects from multiple institutions (http://brain-development.org/ixidataset/). The spatial resolution of all slices is resampled to $0.94 \times 0.94 \times 1.25$ $mm^3$, in-plane matrix size is fixed at $256 \times 256$, and the number of slice ranged of 28–136. This dataset was randomly split into training and validation sets in the ratio of 80%:20%.

**BraTs dataset:** this dataset was collected by MICCAI 2019 grand challenge (https://www.med.upenn.edu/cbica-/brats-2019/). In the current study, 259 subjects with high-grade glioblastoma (GBM) and 76 subjects with low-grade glioblastoma from this dataset were used to test model. As publicly shared, MRI images of this dataset were collected

**Table 1**
Datasets information.

| Parameter | Public Datasets | | In-House Datasets | |
|---|---|---|---|---|
| | IXI (healthy) | BraTs (GBM) | MS | CI |
| Model development | | | | |
| Training | 463 | 0 | 0 | 0 |
| Validation | 115 | 0 | 0 | 0 |
| Testing | 0 | 317 | 40 | 40 |
| held-out validation | 0 | 18 | 0 | 0 |
| Image parameters | | | | |
| Scanner | Diverse | Diverse | Diverse | |
| Resolution ($mm^3$) | $0.94 \times 0.94 \times 1.25$ | $1 \times 1 \times 1$ | $0.5 \times 0.5 \times 7$ | |
| In-plane size | $256 \times 256$ | $240 \times 240$ | Diverse | |
| Z-axis size | 28–136 | 155 | 18–23 | |
| Skull-stripped | No | Yes | No | |

from 19 institutions and co-registered to a standard anatomical template, interpolated to $1 \times 1 \times 1$ mm$^3$ resolution and skull-stripped, and finally cropped to a fixed size of $240 \times 240 \times 155$.

**In-house dataset:** all in-house images were acquired between 2018 and 2021, which were randomly selected from our PACS. Two typical pathologies: multiple sclerosis (MS) and cerebral infarction (CI) were used to test the generalization performance of the model. This dataset contained 80 subjects, 40 cases per lesion. The privacy of the patients was protected by de-identifying images prior analysis. The classification labels in axial-slice level were provided by a radiologist with more than 5 years of experience. Images have different in-plane matrix sizes and the in-plane resolution fixed at 0.5 mm. The number of slice was in the range of 18–23, with 7 mm thickness.

### 2.2. Image preprocessing

T2 volumes from both IXI dataset and in-house dataset were firstly skull-stripped using the HD-BET tool [34]. Then all volumes were rigidly co-registered to the Montreal Neurological Institute (MNI) template (1 mm isotropic version) using the flirt tool of FMRIB software library (FSL 5.0.9) (https://fsl.fmrib.ox.-ac.uk/fsl/fslwiki/FLIRT). Volumes were then cropped to $160 \times 192 \times 96$ to remove excessive blank spaces. Finally, the intensity range was rescaled to [0, 1].

### 2.3. Unsupervised autoencoder training with healthy brain MRIs

The core concept behind UAD architecture is to model the distribution of healthy anatomy of the brain MRIs. Since abnormal structures are never seen during training, it is difficult to be correctly reconstructed from the latent space. As a result, reconstruction errors in the abnormal regions are expected to be higher than normal regions. Based on this assumption, we trained an autoencoder (AE) network on images from normal scans, yielding an encoder and a decoder, and a latent representation of normal anatomical variability. The encoder network $\mathrm{Enc}_\theta(x)$ with parameter $\theta$ maps the input images into a lower dimensional latent representation space $(z)$, from which a decoder $\mathrm{Dec}_\varnothing(z)$ with parameter $\varnothing$ then reconstructed the input as $\hat{x} = \mathrm{Dec}_\varnothing(\mathrm{Enc}_\theta(x))$. Because the abnormality usually presents a continuous change in MRI sequence, we proposed a 3D autoencoder model to capture the spatial information of consecutive layers. An overview of our framework was given in Fig. 1. The encoder consists of a $3 \times 3 \times 3$ convolution with stride 1, followed by a leaky rectified linear unit (Leaky ReLU) and a batch normalization (BatchNorm). And followed by 4 downsampling residual blocks (DRB), each downsampling step has three convolution groups (containing of convolution with stride 2, LeakyReLU, and BatchNorm). The number of channels of the four DRBs are 32, 64, 128 and 256 respectively. Followed a fully connected layer is used to get the

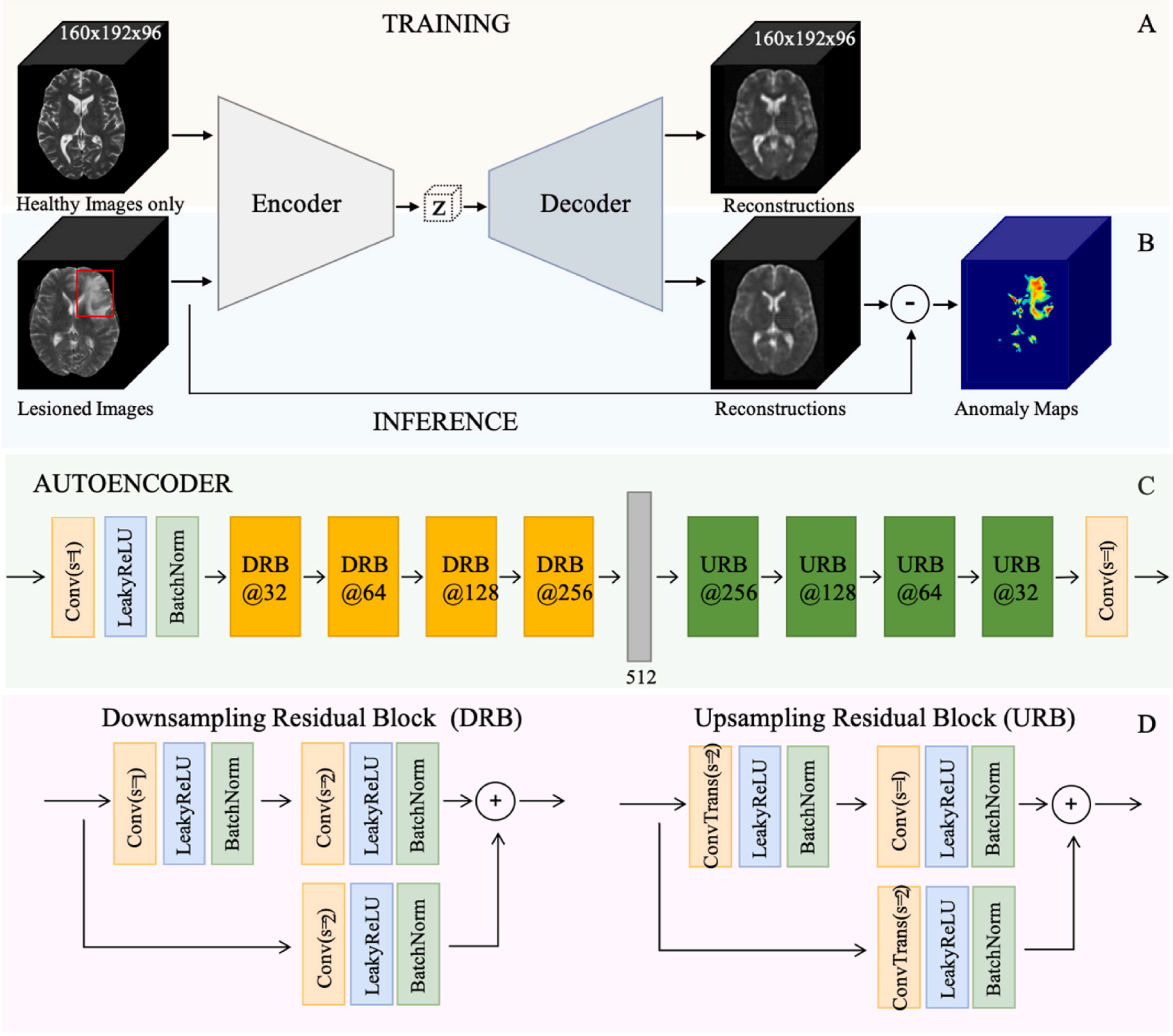


**Fig. 1.** Overview of the proposed autoencoder-based anomaly detection framework: A) Training the autoencoder for modeling the distribution of healthy brain anatomy. B) In the inference phase, anomaly maps reveal lesions by subtracting their reconstruction from the input image. C and D) Details of the proposed autoencoder.

latent space $z$ (default to 512). The decoder is reflective symmetric to the encoder, with downsampling layers replaced by upsampling layers.

In the training stage (see Fig. 1, A), we used training data $x_{healthy}\mathbb{R}^{H\times W\times C}$ sampled from healthy MRI scans, while the multiple abnormalities data $x_{abnormal} \in \mathbb{R}^{H\times W\times C}$ was used in the inference stage. Anomaly maps were calculated by subtracting their reconstruction from the input image. In the case of a normal input image (and under the assumption of a perfect encoder and a perfect decoder), mapping from image space to the latent space via the encoder and subsequent mapping from latent space back to reconstructed image space via the decoder should closely same distribution. However, for lesioned subjects, anomalies are poorly reconstructed once projected to the healthy manifold, and thus yield to high residuals in the anomaly map (see Fig. 1, B). The code for model definition, training and validation has been made publicly available on Github (https://github.com/MAI-Lab-West-China-Hospital/anomaly-detection-of-brain-MR-images).

### 2.4. Model training and optimization

The AE loss was computed as the L2 distance between input volumes ($x$) and their reconstructions ($\widehat{x}$) ($L = \|x - \widehat{x}\|_2$). The size of mini-batches was set to 8, Adam optimizer was used with an initial learning rate of 1e-4 to train the model for 200 epochs. The network parameters were randomly initialized. An L2 regulation on kernel weights with a 1e-5 scale factor was also used to reduce overfitting. All experiments were implemented based on the MONAI (https://monai.io/) and PyTorch framework with dual NVIDIA Tesla V100 with 32 GB memory.

### 2.5. Postprocessing

During testing, we obtained anomaly maps ($V_a$) by computing the residuals between input volumes and their reconstructions ($V_a = |x - \widehat{x}|$). Each anomaly map was shown as a heatmap (Fig. 1, B), in which high residuals depict anomalous structures. We applied some post-processing steps to reduce the number of false positives based on anomaly maps. First, a $5 \times 5 \times 5$ median filter was used to filter out small residuals. Further, we multiplied the anomaly maps with slightly eroded brain masks and truncated by setting a threshold to preserve the high residual parts. Finally, the final anomaly maps have been obtained by applying 3D connected component analysis and removing the smaller connected domains. The results of all models were subject to the same postprocessing.

### 2.6. Performance evaluation and statistical analysis

To quantitatively and qualitatively describe the performance of the proposed method, several evaluation metrics were developed. Specifically, the normality of each axial slice was evaluated by the reconstruction errors. Following [35], we obtained the normality score $s(u)$ of the $u$-th axial slice by normalizing the errors to range [0, 1]. The normality score closer to 0 indicates the axial slice is more likely an abnormal slice. The calculation of the normality score is shown in Appendix E1 (supplement). We used the area under the receiver operating characteristic curve (AUC) as the detection performance metric, calculated from the normality score. However, the ROC curve and AUC are only partially meaningful when used with unbalanced data [36]. Therefore, we also reported the area under the precision-recall curve (AUPRC). To test the sensitivity of detection performance according to whether there is overlap between the ground truth and anomaly map. Further, similar [37], we measured the performance using the best achievable DICE-score ($\lceil DICE \rceil$), which constitutes a theoretical upper-bound to a model's segmentation performance and is obtained via a greedy search for the anomaly maps threshold which yields the highest DICE score on the held-out validation set of 18 patients.

## 3. Experiments and results

### 3.1. Anomalous visualization

Once we trained the model, the normality score can be calculated based on the reconstruction error. Brain MRI axial slices consisting of normal anatomy have a lower error and the abnormal slices have a higher error. The reconstruction error can be calculated for each slice, which enables us to split the error into each axial slice and locate the anomalous region. The normality scores between normal slice and anomalous slice in a single scan were compared in Fig. 2, which shows that the normality score immediately decreases when some anomalies occur.

Visualization examples from three different lesion datasets were shown in Fig. 3. In each lesion, the first row shows the original image, the second row displays the corresponding reconstruction image. The reconstruction error map (anomaly map) is shown as a heatmap in third row, blue represents low error and red represents high. The last row shows the ground truth. If lesions are not within the learned distribution, the model will be difficult to reconstruct this region and it will choose the closest representation that has been learned, which can be considered as a healthy brain representation. Consequently, the lesion region has a higher reconstruction error. Using the threshold to eliminate small errors in anomaly maps can locate the anomalies greatly. It is worth noting that on T2w images, cerebrospinal fluid showed a very high signal and was reconstructed poorly. Therefore, cerebrospinal fluid was also located in the anomaly map.

### 3.2. Anomaly detection performance

In our proposed 3D AE model, the latent vector (z) size determines the degree of compression of the input images. The latent vector can be designed to be dense (dense AE) or spatial (spatial AE). Theoretically, compared with spatial AE, dense AE has higher compression for the scan, but it will also reduce its reconstruction performance. In this section, we conducted several ablation studies to further analyze the effects of different latent vector sizes on anomaly detection performance. We trained several architectures with a varying latent vector sizes. For example, in dense AE, z ranges from 128 to 1024. In spatial AE, the z size was fixed at $10 \times 12 \times 6$ with 1–64 varying channels. The test dataset was all BraTs data. For fair comparison, we drew ROC and PRC curves for each model, and the results were shown in Fig. 4. We observed that all dense AE outperformed spatial AE. Additionally, for dense AE, different latent space sizes have less effect on AUROC, but for spatial AE, the value decreases with the number of channels. It has been observed from Fig. 5, that the 64-channel spatial AE "generalizes" so well that it can also reconstruct anomalies well, leading to the missed anomalies. In conclusion, the model can robustly achieve superior detection performance with the latent space size of 512.

Then we used a model with this optimal latent space to evaluate the anomaly detection performance on the three pathological datasets. We reported the AUC, AUPRC, and Sensitivity at axial slice level (Table 2). On the public BraTs dataset, the proposed model was able to reach AUC value of 0.844, AUPRC of 0.741 and Sensitivity of 0.822. On the two types of lesions of in-house dataset, the AUCs were 0.858 (AUPRC, 0.731, Sensitivity, 0.837) for the multiple sclerosis data and 0.807 (AUPRC, 0.705, Sensitivity, 0.877) for the cerebral infarction data.

### 3.3. Anomaly segmentation performance

In order to evaluate the quality of anomaly segmentation (ie, the accurate coverage of each abnormal voxel), we calculated the $\lceil DICE \rceil$ score on 335 samples from BraTs dataset, which has the ground truth delineated by experts. Since the intensity of anomaly map represents continues normality scores, we computed the $\lceil DICE \rceil$ score by thresholding the anomaly map with a threshold determined by the hold-out

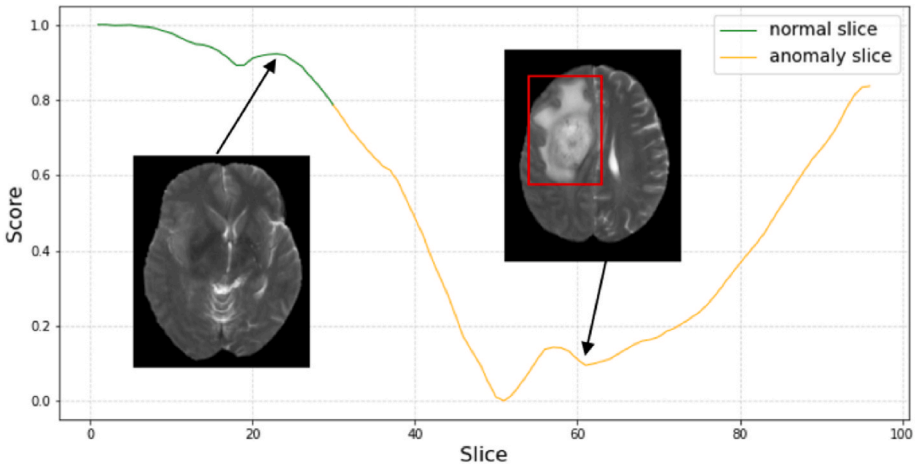


**Fig. 2.** Normality scores of the brain MRI axial slices were obtained by proposed method. The green line represents the normal slices, which hold high normality scores. The orange line represents the anomaly slice, and the score decreases immediately when anomalies occur.

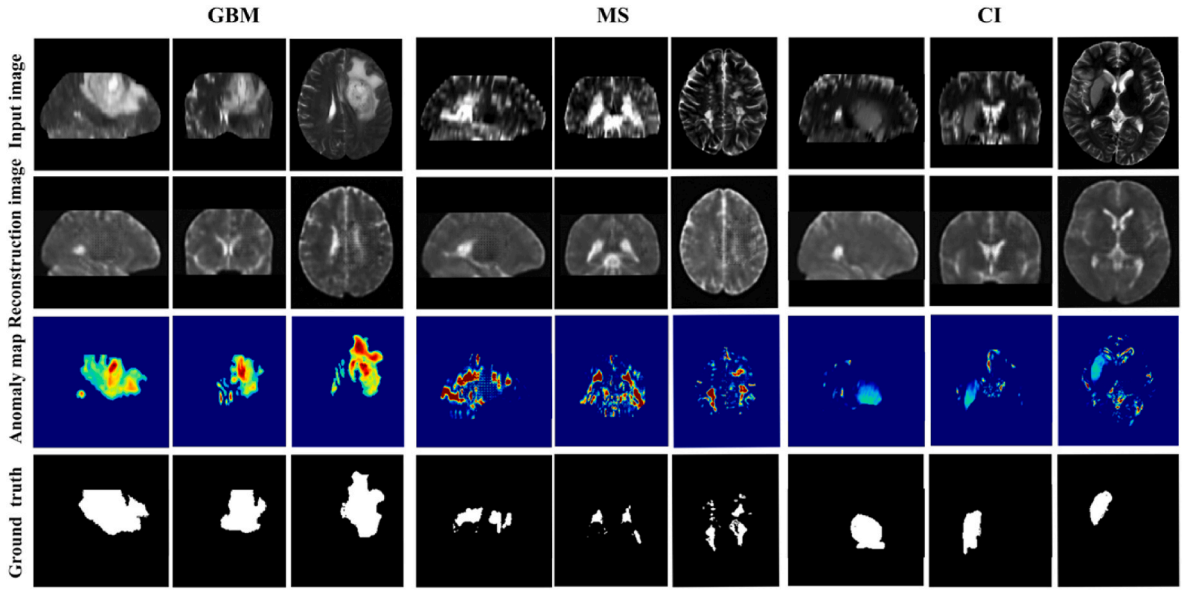


**Fig. 3.** Visualization results of anomalous region localization. First row: Original image. Second row: Reconstruction image. Third row: Anomaly map in voxel-wise level displayed as a heatmap. Fourth row: Ground truth segmentation. From left to right, it shows the detection performance of glioblastoma (GBM), multiple sclerosis (MS) and cerebral infarction (CI).

validation set which contains 18 samples (Table 1). We also compared our proposed method with VAE and GAN, another two popular anomaly detection methods, in terms of lesion segmentation performance. The results were shown in Table 3, our proposed model achieved best ⌈DICE⌉ score of 0.462 and had much higher AUC and AUPRC compared to VAE and f-AnoGAN. Although f-AnoGAN performed well in retinal OCT abnormality detection [28], it falls short of AE and VAE in brain MRI abnormality detection. This is probably due to large structural differences between slices in brain MRI and the fact that encoder and generator of f-AnoGAN are trained in two steps, which makes f-AnoGAN prone to producing incorrect anatomical structures when generating images from latent space.

## 4. Discussion

Glioblastoma, multiple sclerosis, and cerebral infarction are common diseases of the nervous system. In the past, much work has been done to assist in the diagnosis using computers, including conventional computer vision computer vision methods such as level set and graph-cut methods, as well as recently proposed supervised deep learning-based lesion recognition and segmentation methods [8,9,31]. These methods have greatly improved the efficiency and accuracy of disease diagnosis. However, they generalization is limited, meaning that a model trained to segment glioblastoma will have difficulty in recognizing multiple sclerosis or cerebral infarction. In this work, we proposed a 3D autoencoder for UAD in brain MRI. Compared with the previous 2D method, using the entire volume as the input is closer to the behavior of human readers using contextual information by scrolling through slices when reading an image. Given an input, the proposed model first uses the encoder to map the volume to the latent representation and then uses the decoder to obtain its reconstruction. Since the model is only trained to fit the distribution of healthy brains, it can well reconstruct the healthy samples and enlarge the reconstruction residual of the anomalies.

We further investigated the influence of different latent vector sizes on UAD performance. The results showed that all dense AE performed better than the spatial AE. Dense AE have higher compression ratio for the input when compared to that of spatial AE. Consequently, less details are reconstructed and the anomalies were completely ignored, which

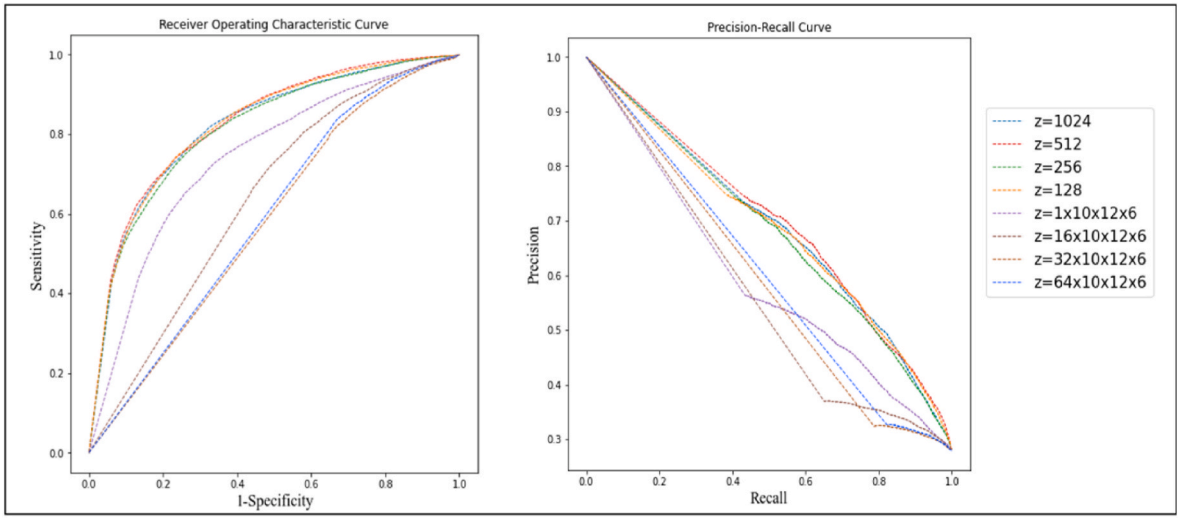


**Fig. 4.** Detection performance for various latent space sizes. Left: ROC curve. Right: Precision-Recall curve.

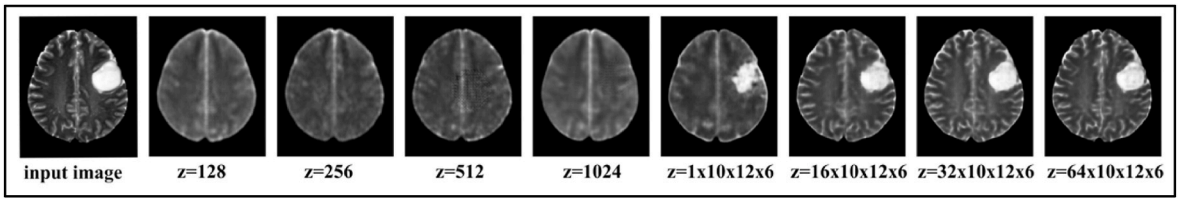


**Fig. 5.** Reconstruction performance for different latent space sizes. The image reconstructed by dense AE is blurry. The reconstruction performance of spatial AE increases with the number of channels, but anomalies are also well reconstructed, which will be detrimental to anomaly detection.

**Table 2**
Anomaly detection performance metrics of three lesions.

| Datasets | AUC | AUPRC | Sensitivity |
|---|---|---|---|
| BraTs (GBM) | 0.844 | 0.741 | 0.822 |
| In-house (MS) | 0.858 | 0.731 | 0.837 |
| In-house (CI) | 0.807 | 0.705 | 0.877 |

**Table 3**
Anomaly detection performance metrics of three lesions.

| Model | AUC | AUPRC | Sensitivity | $\lceil DICE \rceil$ |
|---|---|---|---|---|
| VAE [31] | 0.767 | 0.628 | 0.807 | 0.406 |
| f-AnoGAN [28] | 0.667 | 0.522 | 0.670 | 0.392 |
| AE(Ours) | 0.844 | 0.741 | 0.822 | 0.462 |

makes the anomalous regions have greater reconstruction errors.

We evaluated our best model (z = 512) on three different lesions: glioblastoma, multiple sclerosis and cerebral infarction. Extensive experiments have demonstrated that the proposed model can detect various anomalies. The greater the intensity difference between these lesions and the surrounding normal anatomy, the localization ability of the model is stronger. For slices in a volume, the degree of abnormality was determined by the normality score, the normality score immediately decreases when some anomalies occur. In addition, the preprocessing step that register volumes with arbitrary resolution and rotation to the standard space and unify the resolution further enhances the generalization ability of the model. Segmentation of lesions can be generated from anomaly map by thresholding and the optimal threshold was determined by the highest DICE score achieved on the validation set in the current study. On the BraTs datasets, our UAD model achieved $\lceil DICE \rceil$ score of 0.462. When comparing our method with VAE and f-AnoGAN, we found that images recovered from VAE are blurred, which is an inherent shortcoming of VAE-based models, resulting in poor segmentation performance. f-AnoGAN shows good image "reconstruction" but it is prone to produce incorrect anatomical structures. The segmentation performance of involved unsupervised methods is currently significantly lower than state-of-the-art supervised methods [38,39]. However, the main application field of UAD lies in the distinction of normal and anomalous images and coarsely indicate anomalies. Although the segmentation performance of UAD methods cannot be compared to that of supervised methods, they are still very promising because they can yield models with good generalization performance for a wide range of diseases and do not require manually annotated training samples. Anomalous regions output by UAD can be considered as hotspots, prompting further evaluation by clinical experts, and can also be used as the initialization for any automatic or semi-automatic segmentation algorithm. Anomaly detection can be deployed in during image session and suggested abnormal areas can be used immediately to adjust the acquisition scheme, such as imaging planes or local resolution, or to add modalities that are more valuable for the diagnosis of suspected lesions.

One limitation of our work is that only a single modality is analyzed. Influenced by T2w imaging modality, false positives are often closely associated with regions containing sulci or tissue-cerebrospinal fluid, resulting in lower Dice values. But this effect will not exist on T2-FLAIR modality. In future work, how to extend to multi-sequential MRI data (e.g, contrast-enhanced T1w images, T2-FLAIR, or DWI images) needs further research. In another hand, the most important aspect of such an approach is trust by the medical people into the results. According to Holzinger et al. [40], explainability is one step there and robustness is the second step, and it should be mentioned that explainability and robustness promote reliability and trust in the results and also ensure that humans remain in control. More research needed on how to get

healthcare professionals to truly trust AI.

In summary, we developed an anomaly detection model only using normal brain MR images, which demonstrated the ability to detect a wide range of anomalies on two independent test datasets. Since the model training doesn't need labeled data, it overcomes the typical obstacles occurred in training supervised deep learning methods. This method has the potential to be generalized to other parts and modalities and essentially supported the diagnostic workflow in radiology as an automated tool for computer-aided image analysis.

## Data availability statement

This in-house dataset was currently the private property of West China Hospital, However, the de-identified images would be available upon reasonable request to corresponding author.

## Declaration of competing interest

All authors disclosed no relevant relationships.

## Acknowledgments

This work was supported by China National Key R&D Program (No. 2020AAA0105000 and 2020AAA0105005), National Natural Science Foundation of China (81974278) and Young Elite Scientists Sponsorship Program (YESS20160060) by China Association for Science and Technology.

## Appendix E1. Normality Score

After the image is reconstructed by the autoencoder, the reconstruction error is expressed as the absolute error between the input image and its reconstruction. Abnormalities are more likely to stand out, as the trained model does not have the necessary information to accurately reconstruct or predict anomalies. Reconstruction errors also called anomaly map, which represents voxel-wise error of the whole scan. For a 3D brain MRI scan, the normality score can be used to quantitatively evaluate the degree of abnormality of each slice at the slice level, which is computed from the reconstruction errors. The normality score sums the reconstruction errors of each slice and normalizes them to [0, 1], as different scans may have different notions of abnormality. We compute the normality score $s(u)$ of a slice $u$ as follows:

$$s(u) = 1 - \frac{e(u) - min_u e(u)}{max_u e(u) - min_u e(u)}$$

Where $u$ is the $u$-*th* slice of the whole scan and $e(u)$ is the reconstruction error of that slice. The slices containing normal anatomy have a higher normality score since they are similar to the data used to train the model, while slices containing abnormal structures have a lower normality score.

## References

[1] C. Jean-Quartier, F. Jeanquartier, A. Holzinger, Open data for differential network analysis in glioma, Int. J. Mol. Sci. 21 (2020) 547, https://doi.org/10.3390/ijms21020547.

[2] A. Kopkova, J. Sana, T. Machackova, M. Vecera, L. Radova, K. Trachtova, V. Vybihal, M. Smrcka, T. Kazda, O. Slaby, P. Fadrus, Cerebrospinal fluid MicroRNA signatures as diagnostic biomarkers in brain tumors, Cancers 11 (2019) 1546, https://doi.org/10.3390/cancers11101546.

[3] T.J. Murray, Diagnosis and treatment of multiple sclerosis, BMJ 332 (2006) 525–527, https://doi.org/10.1136/bmj.332.7540.525.

[4] W.J. Brownlee, T.A. Hardy, F. Fazekas, D.H. Miller, Diagnosis of multiple sclerosis: progress and challenges, Lancet 389 (2017) 1336–1346, https://doi.org/10.1016/S0140-6736(16)30959-X.

[5] T.H. Shin, D.Y. Lee, S. Basith, B. Manavalan, M.J. Paik, I. Rybinnik, M. M. Mouradian, J.H. Ahn, G. Lee, Metabolome changes in cerebral ischemia, Cells 9 (2020) 1630, https://doi.org/10.3390/cells9071630.

[6] M.E. Tschuchnig, M. Gadermayr, Anomaly detection in medical imaging - A mini review, Data Science–Analytics and Applications (2022) 33–38, https://doi.org/10.1007/978-3-658-36295-9_5, 2022.

[7] G. Pang, C. Shen, L. Cao, A. van den Hengel, Deep learning for anomaly detection, ACM Comput. Surv. 54 (2022) 1–38, https://doi.org/10.1145/3439950.

[8] S. Taheri, S.H. Ong, V.F.H. Chong, Level-set segmentation of brain tumors using a threshold-based speed function, Image Vis Comput. 28 (2010) 26–37, https://doi.org/10.1016/j.imavis.2009.04.005.

[9] D. García-Lorenzo, J. Lecoeur, D.L. Arnold, D.L. Collins, C. Barillot, Multiple Sclerosis Lesion Segmentation Using an Automatic Multimodal Graph Cuts, 2009, pp. 584–591, https://doi.org/10.1007/978-3-642-04271-3_71.

[10] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, H. Larochelle, Brain tumor segmentation with deep neural networks, Med. Image Anal. 35 (2017) 18–31, https://doi.org/10.1016/j.media.2016.05.004.

[11] M. Liu, F. Li, H. Yan, K. Wang, Y. Ma, L. Shen, M. Xu, A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in Alzheimer's disease, Neuroimage 208 (2020), 116459, https://doi.org/10.1016/j.neuroimage.2019.116459.

[12] A. Alijamaat, A. NikravanShalmani, P. Bayat, Multiple sclerosis lesion segmentation from brain MRI using U-Net based on wavelet pooling, Int. J. Comput. Assist. Radiol. Surg. 16 (2021) 1459–1467, https://doi.org/10.1007/s11548-021-02327-y.

[13] B.H. Menze, A. Jakab, S. Bauer, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imag. 34 (2015) 1993, https://doi.org/10.1109/TMI.2014.2377694. –2024.

[14] O. Maier, B.H. Menze, J. von der Gablentz, et al., Isles 2015 - a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI, Med. Image Anal. 35 (2017) 250–269, https://doi.org/10.1016/j.media.2016.07.009.

[15] S. Wang, Y. Cong, H. Zhu, X. Chen, L. Qu, H. Fan, Q. Zhang, M. Liu, Multi-scale context-guided deep network for automated lesion segmentation with endoscopy images of gastrointestinal tract, IEEE J Biomed Health Inform 25 (2021) 514–525, https://doi.org/10.1109/JBHI.2020.2997760.

[16] K. Hu, L. Zhao, S. Feng, S. Zhang, Q. Zhou, X. Gao, Y. Guo, Colorectal polyp region extraction using saliency detection network with neutrosophic enhancement, Comput. Biol. Med. 147 (2022), 105760, https://doi.org/10.1016/j.compbiomed.2022.105760.

[17] H. Su, D. Zhao, H. Elmannai, A.A. Heidari, S. Bourouis, Z. Wu, Z. Cai, W. Gui, M. Chen, Multilevel threshold image segmentation for COVID-19 chest radiography: a framework using horizontal and vertical multiverse optimization, Comput. Biol. Med. 146 (2022), 105618, https://doi.org/10.1016/j.compbiomed.2022.105618.

[18] A. Qi, D. Zhao, F. Yu, A.A. Heidari, Z. Wu, Z. Cai, F. Alenezi, R.F. Mansour, H. Chen, M. Chen, Directional mutation and crossover boosted ant colony optimization with application to COVID-19 X-ray image segmentation, Comput. Biol. Med. 148 (2022), 105810, https://doi.org/10.1016/j.compbiomed.2022.105810.

[19] Y. Tian, G. Pang, F. Liu, Y. Chen, S.H. Shin, J.W. Verjans, R. Singh, G. Carneiro, Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2021, pp. 128–140, https://doi.org/10.1007/978-3-030-87240-3_13.

[20] Yunqiang Chen, Sean Zhou Xiang, T.S. Huang, One-class SVM for learning in image retrieval, in: Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205), IEEE, n.d.: pp. 34–37. https://doi.org/10.1109/ICIP.2001.958946.

[21] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S.A. Siddiqui, A. Binder, E. Müller, M. Kloft, Deep one-class classification, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, PMLR, 2018, pp. 4393–4402, in: https://proceedings.mlr.press/v80/ruff18a.html.

[22] A. Zimek, E. Schubert, H.-P. Kriegel, A survey on unsupervised outlier detection in high-dimensional numerical data, Stat. Anal. Data Min. 5 (2012) 363–387, https://doi.org/10.1002/sam.11161.

[23] L. Xiong, B. Póczos, J. Schneider, Group anomaly detection using flexible genre models, in: J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, K.Q. Weinberger (Eds.), Adv Neural Inf Process Syst, Curran Associates, Inc., 2011, in: https://proceedings.neurips.cc/paper/2011/file/eaae339c4d89fc102edd9dbdb6a28915-Paper.pdf.

[24] H. Liu, X. Xu, E. Li, S. Zhang, X. Li, Anomaly detection with representative neighbors, IEEE Transact. Neural Networks Learn. Syst. (2021) 1–11, https://doi.org/10.1109/TNNLS.2021.3109898.
[25] D. Gong, L. Liu, V. Le, B. Saha, M.R. Mansour, S. Venkatesh, A. van den Hengel, Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection, 2019.
[26] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, X.-S. Hua, Spatio-temporal AutoEncoder for video anomaly detection, in: Proceedings of the 25th ACM International Conference on Multimedia, ACM, New York, NY, USA, 2017, pp. 1933–1941, https://doi.org/10.1145/3123266.3123451.
[27] C. Zhou, R.C. Paffenroth, Anomaly detection with robust deep autoencoders, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2017, pp. 665–674, https://doi.org/10.1145/3097983.3098052.
[28] T. Schlegl, P. Seeböck, S.M. Waldstein, G. Langs, U. Schmidt-Erfurth, f-AnoGAN, Fast unsupervised anomaly detection with generative adversarial networks, Med. Image Anal. 54 (2019) 30–44, https://doi.org/10.1016/j.media.2019.01.010.
[29] K.M. van Hespen, J.J.M. Zwanenburg, J.W. Dankbaar, M.I. Geerlings, J. Hendrikse, H.J. Kuijf, An anomaly detection approach to identify chronic brain infarcts on MRI, Sci. Rep. 11 (2021) 7714, https://doi.org/10.1038/s41598-021-87013-4.
[30] B. Lambert, M. Louis, S. Doyle, F. Forbes, M. Dojat, A. Tucholka, Leveraging 3d information in unsupervised brain mri segmentation, in: Proceedings - International Symposium on Biomedical Imaging, IEEE, 2021, pp. 187–190, https://doi.org/10.1109/ISBI48211.2021.9433894.
[31] C. Baur, S. Denner, B. Wiestler, N. Navab, S. Albarqouni, Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study, Med. Image Anal. 69 (2021), 101952, https://doi.org/10.1016/j.media.2020.101952.
[32] C. Baur, B. Wiestler, M. Muehlau, C. Zimmer, N. Navab, S. Albarqouni, Modeling healthy anatomy with artificial intelligence for unsupervised anomaly detection in brain MRI, Radiol Artif Intell 3 (2021), e190169, https://doi.org/10.1148/ryai.2021190169.
[33] Y. Chen, H. Zhang, Y. Wang, Y. Yang, X. Zhou, Q.M.J. Wu, Mama Net, Multi-scale attention memory autoencoder network for anomaly detection, IEEE Trans. Med. Imag. 40 (2021) 1032–1041, https://doi.org/10.1109/TMI.2020.3045295.
[34] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H. Schlemmer, S. Heiland, W. Wick, M. Bendszus, K.H. Maier-Hein, P. Kickingereder, Automated brain extraction of multisequence MRI using artificial neural networks, Hum. Brain Mapp. 40 (2019) 4952–4964, https://doi.org/10.1002/hbm.24750.
[35] M. Hasan, J. Choi, J. Neumann, A.K. Roy-Chowdhury, L.S. Davis, Learning Temporal Regularity in Video Sequences, 2016. http://arxiv.org/abs/1604.04574.
[36] A.M. Carrington, P.W. Fieguth, H. Qazi, A. Holzinger, H.H. Chen, F. Mayr, D. G. Manuel, A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms, BMC Med. Inf. Decis. Making 20 (2020) 4, https://doi.org/10.1186/s12911-019-1014-6.
[37] W.H.L. Pinaya, P.-D. Tudosiu, R. Gray, G. Rees, P. Nachev, S. Ourselin, M. J. Cardoso, Unsupervised Brain Anomaly Detection and Segmentation with Transformers, 2021. http://arxiv.org/abs/2102.11650.
[38] Z. Jiang, C. Ding, M. Liu, D. Tao, Two-stage cascaded U-net: 1st place solution to BraTS challenge 2019, Segmentation Task (2020) 231–241, https://doi.org/10.1007/978-3-030-46640-4_22.
[39] M. Islam, V.S. Vibashan, V.J.M. Jose, N. Wijethilake, U. Utkarsh, H. Ren, Brain Tumor Segmentation and Survival Prediction Using 3D Attention UNet, 2020, https://doi.org/10.1007/978-3-030-46640-4_25, 262–272.
[40] A. Holzinger, Frontier: AI We Can Really Trust (2021), https://doi.org/10.1007/978-3-030-93736-2_33, 427–440.